The Dissertation Committee for Cynthia Esperanza Lima Gonzalez Certifies that this is the approved version of the following dissertation:

The role of language and culture in large-scale assessment: A study of the 2009 Texas Assessment of Knowledge and Skills

Committee:

Guadalupe Carmona, Supervisor

Jill A. Marshall

Walter M. Stroup

Maria E. Franquiz

Daniel A. Powers

The role of language and culture in large-scale assessment: A study of the 2009 Texas Assessment of Knowledge and Skills

by

Cynthia Esperanza Lima Gonzalez, B.S., M.A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 2013

The role of language and culture in large-scale assessment: A study of the 2009 Texas Assessment of Knowledge and Skills

Cynthia Esperanza Lima Gonzalez, Ph. D.

The University of Texas at Austin, 2013

Supervisor: Guadalupe Carmona

The inclusion of all students in large-scale assessment mandated by the No Child Left Behind (2003) requires that these large-scale assessments be developed to allow *all* students to show what they know, and that the results are *comparable* and *equitable* across diverse cultural and linguistic populations. This study examined the validity of the 5th grade 2009 Science Texas Assessment of Knowledge and Skills (TAKS) for diverse cultural and linguistic groups. The student groups considered for this study were selected based on all the possible combinations of three variables: ethnicity –White and Hispanic, test language –English and Spanish, and Limited English Proficiency (LEP) classification.

Validity was assessed at the item and construct levels, and was analyzed from a psychometric, cultural and linguistic stance.

At the item level, Differential Item Function (DIF) was conducted using the Mantel-Haenszel procedure. The presence of biased items was revealed for all pairwise group comparisons; with a high number of DIF items between groups which differed in English

proficiency (approximately 50% of the test items), and a low number of DIF items between groups which only differ in ethnicity (approximately 15% of the test items). However, an analysis of the Item Characteristic Curves (ICCs), revealed that items classified by the Mantel-Haenszel procedure as advantaging the LEP groups, did so for students at low proficiency levels; while the advantage at high proficiency levels was for non-LEP groups.

At the construct level, the structure of the English version of the TAKS was compared across three student groups using Confirmatory Factor Analysis with Multiple Groups. The hypothesized structure based on the TAKS blueprint, was rejected for the Group conformed by White, non-LEP students ($MLM\chi^2_{[734]} = 1042.110$; CFI= 0.845; RMSEA= 0.020); but, it was a good fit for Hispanic, non-LEP ($MLM\chi^2_{[734]} = 819.356$; CFI= 0.980; RMSEA= 0.011) and LEP ($MLM\chi^2_{[734]} = 805.124$; CFI= 0.985; RMSEA= 0.010) Groups.

The results obtained from this study call to reinterpret the achievement gap observed in TAKS scores between the populations considered, and highlight the need for further development of guidelines that can better help to develop fair large-scale tests for all students.

Table of Contents

List of Tables

List of Figures

Chapter I: Introduction

THE ACHIEVEMENT GAP

The achievement gap is one of the most discussed issues in education. It usually refers to differences in standardized test scores between White and African American and White and Hispanic students (Ladson-Billings, 2006). The gap in achievement has been widely reported and persistent, even though it narrowed in the 1970s and 1980s (Johnson, 2002; Lee, 2002); and it has raised awareness among educators regarding the way students are being taught, the role of culture and language in learning, and how the curriculum should be modified to teach all students (Johnson, 2002; Lee & Buxton, 2010; Singham, 2003). To better understand the term *achievement gap*, I briefly discuss the definition of achievement and the way in which it is measured.

Achievement is generally used to refer to what students are able to do. Through the years the expectations of what students should be able to do in Science have changed. Expectations also vary depending on each state mandated curriculum –e.g. Texas curriculum for fifth grade Science differs from the California curriculum for the same grade level. Although these expectations may differ depending on the state, the National Research Council (2007) provides the following learning goals for K- 8 students that provide a conventionally accepted definition of science proficiency:

1. know, use and interpret scientific explanations of the natural world;
2. generate and evaluate scientific evidence and explanations;
3. understand the nature and development of scientific knowledge; and

4. participate productively in scientific practices and discourses. (p. 2)

These learning goals incorporate both scientific knowledge and skills -since both are considered to be intertwined, and needed for students to be educated citizens and active members of the society (National Research Council [NRC], 2007).

Texas fifth grade curriculum is divided in four content areas: (a) Nature of Science, (b) Life sciences, (c) Physical Science, and (d)Earth Science, (Texas Education Agency [TEA], 2004). The Texas Essential Knowledge and Skills (TEKS) describe the knowledge and skills that students should master related to the four content areas. Student achievement is measured using a large-scale assessment. The Texas Assessment of Knowledge and Skills (TAKS) is the large-scale assessment used from 2003 to 2011 in the state of Texas to assess the degree to which students learned the state mandated curriculum following a more comprehensive approach than the previous tests –e.g. Texas Assessment of Academic Skills (TAAS). The mandated large-scale assessment is implemented state wide, and test scores are monitored for different student groups based on ethnicity and English proficiency, among other student information. Large-scale assessment scores are used to follow trends in student achievement.

When comparing achievement between student groups using a large-scale instrument, it is crucial to keep in mind what is being assessed by the test. For instance, Darling-Hammond et al. (2008) address the fact that large-scale assessments such as the National Assessment for Educational Progress (NAEP) and the Trends in International Mathematics and Science Study (TIMSS) evaluate whether students learned what they

were taught, while PISA is intended to evaluate students' ability to apply what they learned.

In the case of Texas, achievement gaps constitute differences in attainment scores of an assessment intended to measure whether students have learned the state mandated curriculum.

ASSESSING HISPANIC AND ENGLISH LANGUAGE LEARNERS STUDENTS

According to the U.S. Department of Education (2006), Limited English Proficient (LEP) students are the fastest-growing population. One of every four students is expected to be LEP by 2025. In 2009, the number of students enrolled in Texas Public Schools from early education to 12, was 4,749,571, from which 48% were Hispanic, and 16% were in a bilingual or English as second language program (TEA, 2010). In 2012 the total student population in Texas was 4,998,579 students; the Hispanic population constituted 50% of the total student population, and students in a bilingual or English as a second language program comprised 16% of the total (TEA, 2012). Both, Hispanic and LEP student populations are two of the groups identified to score lower in large-scale assessments such as NAEP than their mainstream peers. Studies reporting on the White-Hispanic gap in Texas include the one conducted by Klein, Hamilton, McCaffrey and Stecher (2000). They compared fourth graders Math scores in the TAAS over a four year period (1994-1998), finding that the Hispanic-White achievement gap narrowed during these years. Similar findings were reported by Linton and Kester (2003) for eight grade students' Math TAAS scores over a four year period (1996-2000).

The disparities of Hispanic and English Language Learner (ELL) students' scores in comparison to White students have produced important changes in education policy such as the ones addressed by No Child Left Behind (2003). One of the actions taken under this policy was the inclusion of ELLs in large-scale assessments programs to make sure that they are being taught effectively (U. S. Department of Education, 2006). Other efforts to serve ELLs include improvement of content assessments, providing students with translations of the English test to their native language, and provision of testing accommodations (e.g. bilingual dictionary, glossaries, etc.). The main objective of these policies is to measure what "LEP students know and what they have learned in all subjects so instructional decision can be based on valid and reliable data" (U.S. Department of Education, 2006, para. 4). Initially these efforts were concentrated on the assessment of literacy and numeracy, but recently, Science has been added to the mandated content assessed. Nevertheless, still not much is known about the extent to which providing ELL students with accommodations yields more valid test scores, or whether students are better able to demonstrate what they know when they are tested in their native language (Abedi, 2011; Abedi, Courtney, Mirocha, Leon, & Goldberg, 2005; Lee & Buxton, 2010). Though it is recognized that linguistic and cultural factors influence students' responses to items, and that assessments are "inevitable cultural products" (Lee & Luykx, 2006, p. 94; Solano-Flores et al., 2001), these factors are likely to contribute to the achievement gap. For instance, Abedi et al. (2005) conducted a study in which they examined the effectiveness of four different accommodations, finding that providing students with an English dictionary was among the more effective

accommodations for ELL students in Grade 4. Additionally, they reported that the use of certain accommodations increased the performance of ELL and reduced the performance gap between ELL and non-ELL; suggesting that language is a factor that influences students' interpretation of the items leading to lower test scores.

RATIONALE

Large-scale assessments that are able to provide valid and valuable information about students' knowledge that can be used to improve instruction is one of the assessment needs underscored by the NRC, in its 2001 report *Knowing what students know*: *The science and design of educational assessment*. This need is true for all students, and it becomes imperative in addressing the achievement gaps, the growing population of ELLs, and the existing threats to validity (e.g. content bias) when using large-scale assessments to evaluate ELLs (U.S. Department of Education, 2006; Lee & Luykx, 2006; Solano-Flores, 2011).

Although TEA provides a Spanish version of the large-scale assessment to 3-5 grade students, it is still not known the extent to which this accommodation controls for test language and/or cultural biases. To date, there is no consensus of the best approach to assess students from diverse linguistic and cultural backgrounds (Abedi, Hofstetter, & Lord, 2004; Lee & Luykx, 2006; Solano-Flores & Trumbull, 2008).

PURPOSE

The purpose of this dissertation is to provide validity evidence of fifth grade Hispanic and ELLs 2009 Science TAKS scores to re-interpret the differences in students' scores and deepen the understanding of the role of large-scale assessment on measuring

differences in student achievement. This study will also shed light on the extent on which the transadaptation process followed to generate the TAKS Spanish version is useful to eliminate linguistic and cultural biases. The purpose of this study is not to determine *the* validity of the fifth grade 2009 Science TAKS scores, but rather to initiate a process of validation for which more evidence should be generated.

RESEARCH QUESTIONS

The generation process of the validity evidence of the fifth grade 2009 Science TAKS scores will be guided by the following research questions:

a. Are there differences in science scores between fifth grade students from different ethnic and linguistic backgrounds who answered the English or Spanish versions of the 2009 Science Texas Assessment of Knowledge and Skills?

b. Is the probability of endorsing[1] the 2009 fifth grade Science Texas Assessment of Knowledge and Skills items the same for students who answered the English or Spanish version of the test?

c. Are the constructs measured by the 2009 fifth grade Science Texas Assessment of Knowledge and Skills equivalent across students from different ethnic and linguistic backgrounds?

LIMITATIONS

One of the main problems in conducting studies with students from diverse linguistic and cultural backgrounds is the characterization of student groups. Even when

[1] The term endorsing used in research question b and throughout the dissertation, refers to answering an item correctly.

language, culture and ethnicity are terms widely used to characterize students, most of the times they stereotype populations (Lee & Luykx, 2006). According to Lee and Buxton (2010), "ethnicity is generally used to represent membership in a social group with shared history, sense of identity, geography, and cultural roots" (p. 12). In this sense, ethnicity is used to refer a group of people who are very similar. However, in the case of this study, the student data collected by TEA (2009a), only recognizes five ethnic groups: (a) American Indian or Alaskan Native, (b) Asian or Pacific Islander, (c) African American, (d) Hispanic, and (e) White. This definition limits student characterization by naming "Hispanic" a heterogeneous group that come from a variety of countries, which might speak diverse Spanish dialects and might have important cultural differences in terms of each of the elements that characterize people considered from the same ethnicity -e.g. country of origin and shared history, as described by Lee and Buxton (2010). Consequently, the use of such a broad characterization, poses some limitations on our understanding of the populations that are more impacted by cultural and linguistic factors embedded in the fifth grade 2009 Science TAKS assessment.

Chapter II: Literature Review

The interest to compare assessment results of students from different backgrounds, languages and culture has continuously been growing with the use of international tests (Hambleton, 2005). In the United States the interest in the performance of students from different cultural and linguistic backgrounds emerged from the accountability system requiring that *all* students, including ELLs, achieve given academic standards. Large-scale assessments have been the vehicle to ensure that these educational goals are met. The inclusion of ELLs challenges the assumption that large-scale assessment allows *all* students to demonstrate what they know. Research on the effects of ELLs' language and culture in their responses to science items/tasks show that large-scale assessments might not reflect what students know (Lee & Buxton, 2010; Luykx, et al, 2007; Solano-Flores & Nelson-Barber, 2001).

Research related to the assessment of ELLs face different challenges. The first challenge is related to threats to the validity of large-scale scores. The second one is finding a common definition of who is an ELL. There are various ways in which the term ELL is used in research. In the following sections I review the different definitions of ELL used in research, and provide an operational definition that was used for this study. I then highlight the main threats to the validity of ELLs' large-scale test scores.

IDENTIFYING ENGLISH LANGUAGE LEARNERS

English Language Learners are defined by Lacelle-Peterson and Rivera (1994), as "students whose first language is not English, and encompasses both students who are

just beginning to learn English (often referred to as "limited English proficient" or "LEP") and those who have already developed considerable proficiency" (p. 55). Despite the fact that ELLs are most of the times characterized as a group of students who are developing English proficiency, this definition includes students from different cultural, linguistic and family backgrounds that make this group highly diverse.

The term "ELL" refers to students whose native language is not English and entails the adverse impact of their low English proficiency level in their academic achievement (Noble et al., 2012; Solano-Flores & Gustafson, 2013).

In this study I used data that was collected and provided by TEA, thus, I used the "ELL" definition provided by TEA: "Student of limited English proficiency means a student whose primary language is other than English and whose English language skills are such that the student has difficulty performing ordinary classwork in English" (Texas Education Code §29.052, p.18). TEA uses the terms Limited English Proficiency (LEP) and ELL interchangeably.

THREATS TO THE VALIDITY OF LARGE-SCALE SCIENCE TEST SCORES

Equitable assessment for ELLs is conceptualized by Lacelle-Peterson and Rivera (1994) as the one that "allows students to show their knowledge, skills and abilities, through the medium of the language or languages in which the material was taught" (p. 66). Equity from a psychometric perspective translates in using the same test for all the students and the same testing conditions. Thus, a test that should not contain systematic errors "in how a test measures for members of a particular subgroup" (Camilli &

Shepard, 1994, p. 8), meaning that existing differences in test scores between groups are real differences in achievement.

The main approaches to generate valid assessments for ELLs in the U.S. have involved the use of an English version of the original tests with students from diverse cultures but adjusting it for item bias, or using adapted versions (Solano-Flores & Nelson-Barber, 2001). Both approaches are based on the use of a test that is originally developed with the purpose to assess white, middle class, English speaker students (Butler & Stevens, 2001; Laosa, 1977). Consequently, it is important to discuss item and construct bias. By *item bias* I refer to any threats that affect items individually, including poor translations or wording. I use c*onstruct bias* to refer to the presence of any construct in the assessment other than the one intended to be measured.

The following pages provide a review of these two sources of bias (Fig.1), followed by a section on two approaches that have been documented in the literature to provide equitable assessment to ELLs: accommodations and the notion of cultural validity (Solano-Flores & Nelson-Barber, 2001).

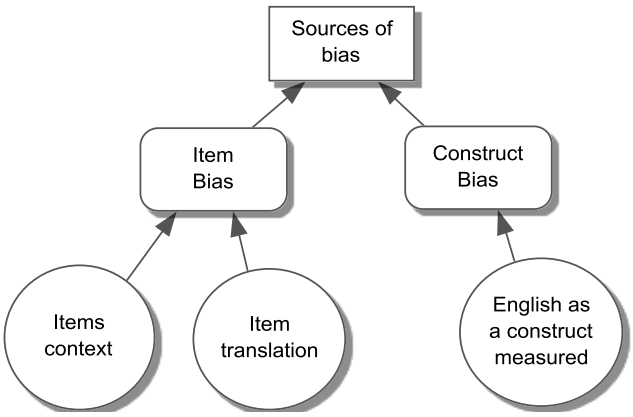


Figure 1: Sources of bias for ELL large-scale assessment

Item bias

> Tests are cultural artifacts… they are written in the language… used by those who develop them; their content is a reflection of the skills, competencies forms of knowledge, and communication styles valued by a society –or the influential groups of that society; and they assume among test-takers full familiarity with the contexts used to frame problems, the ways in which questions are worded, and the expected ways to answer those questions (Solano-Flores, 2011, p. 1).

Most of the large-scale assessments used across diverse language populations in the U.S. are either created to assess English native speakers, or are translations of such tests (Butler & Stevens, 2001). Thus, standardized tests are likely to reflect the values, language and culture of the dominant population (i.e. white, middle class, English-speaking students) (Laosa, 1977). When the same standardized test is used across populations that differ in language and culture, items might not function in the same way for all student subpopulations. This is, items might not measure the content intended in

the same way for all students, since other factors can interfere with the measurement. The sources of item bias reported in the literature include (1) items eliciting knowledge other than the intended and (2) poorly translated items (Garcia & Pearson, 1994; Geisinger, 1994; Hambleton, 2005; van de Vijver & Poortinga, 2005).

Items context and expected thinking

According to Solano-Flores and Nelson-Barber (2001), the action of answering a standardized test and knowing the behaviors expected are also rooted in culture. Consequently students who are not accustomed to the assessment culture might encounter difficulties in answering the test. These difficulties include figuring out ways to answer items and becoming familiar with the format and discourse used to write the items. For instance, the situation(s) provided in the items' stem to contextualize the problems are used under the assumption that they will elicit specific thinking and behaviors in students (Haladyna, Downing, & Rodriguez, 2002), but this assumption might lead to misinterpretation of ELLs' thinking and knowledge. When the same test is used across cultures, some students might not be familiar with the contexts or situations introduced, and the assumptions of the behaviors and thinking that an item should elicit can change. Thus, responses that might be considered "wrong" are likely to be aligned to the personal interpretations of students and make sense in a cultural context which is likely to be different than the one assumed during the test construction. Examples of this are provided and discussed in the literature.

Luykx et al. (2007) found that students' responses to items were based on their beliefs, home culture or the priority of their parents' admonitions. Answers to the

question: “You hear the weatherman say it is 93° Fahrenheit this afternoon. Do you think you will need a sweater if you go outside to play? Explain your answer.” (Luykx et al., 2007, p. 912), included: “Yes, because you always take a sweater when you go outside” (Luykx et al., 2007, p. 912); which was considered by the researchers to reflect the priority of parent’s admonitions. Perhaps one of the most interesting examples of how the home culture influences students’ responses to items are the answers to a question about how long a kid could play outside if it is now 4:00 pm and the kid had to get back home to dinner at 6:00 pm (Luykx et al., 2007, p. 909). According to the researchers, Haitian students were puzzled by the idea of eating dinner at 6 pm, as a similar word for dinner in Haitian is used to refer to the meal eaten at midday. Other students answered to this problem by explaining they would do their homework before going out to play, or apologizing for not knowing the answer.

In another study, Solano-Flores and Nelson-Barber (2001) report how a student relied on her personal experience to answer a question from the National Assessment of Educational Progress (NAEP). The exercise showed a picture of two mountains, one flat and round, and the second one peaked, and asked students to circle the picture that showed how mountains are shaped nowadays. The interviewed student didn’t remember having learned about mountains in school, but she recalled looking at some peaked mountains. The reliance in her personal experience with geologically new mountains along with the lack of knowledge about weathering, led her to provide a “wrong” answer to this question. According to the researchers, this item might privilege students with

experiences with flat and round mountains over students with experiences with peaked mountains (Solano-Flores & Nelson-Barber, 2001).

Another issue relevant to ELLs testing is assuming that all students taking the same test have had the same classes and their acquired knowledge is the same. This assumption is challenged when testing ELLs, especially students who had migrated recently to the US. Teachers participating in the study conducted by Solano-Flores, Sexton, Lara and Navarrete (2001), discussed how probability is not taught formally in China during the first nine years of instruction, therefore, it could be expected that Chinese students would perform low in problems assessing students' knowledge of probability.

This body of research, suggest that students are likely to rely heavily on their previous experiences to answer science items. Thus, their answers are likely to be shaped by "cultural practices, norms and beliefs characteristic of home environment" (Luykx et al., 2007, p. 910). The ways in which knowledge and contexts are conceptualized among cultures vary significantly, even among populations considered to belong to the same cultural group living a few miles apart (Solano-Flores & Nelson-Barber, 2001). The impact of culture in students' science achievement is not only particular to situations in which ELLs are tested in English, but also when they are provided with translated tests, because in test translations construct equivalence is difficult to be maintained (Abedi et al., 2004).

Item translation

Translating items or tests to different languages has become popular due to the continued use of large-scale assessments across countries. In the case of Texas, TEA

provides a Spanish version of the standardized test to Spanish native speakers. Even when the use of test translation is common, there is still ongoing discussion regarding the translation process that yields better results in terms of assessing the same content without varying the level of difficulty across cultures. Moreover, if items are not properly adapted, poor translations can become a source of measurement error (Solano-Flores & Gustafson, 2013). The translation designs commonly used are double translation, back-translation and transadaptation.

The Program for International Student Assessment (PISA) uses a *double translation* from two different languages. During this process the test is translated from two sources to a target language and then reconciled by a third translator (Grisay, 2003).

The *back-translation* design requires two translators to adapt a test from its original language to the target language. Different translators then adapt the test in the target language to the original source language. Both versions are compared for equivalence. Even when this process provides evidence of the fidelity of the translated version, it is not likely to support its valid use (Hambleton, 2005). Brislin and Freimanis (2001) consider that back-translation is a good option when the researcher is not familiar with the target language and has to rely on the translator.

When it comes to standardized tests for students in the U.S., Pearson developed a *transadaptation* guide to provide "high-quality education assessments" available for different linguistic populations in response to the inclusion of ELLs to the accountability system (Zucker, Miska, Alaniz, & Guzmán, 2005). This process consists on adapting the context of the original item (English version), so it is culturally sound for the target

population, and then translating the item from English to the target language. A second translator evaluates the final item for discrepancies in the translation.

One of the challenges of adapting tests is having a set of guidelines for the process (Hambleton, 2005; Solano-Flores, Backhoff, & Contreras-Niño, 2009). The most known and advanced work regarding test adaptation guidelines is the work conducted by the International Test Commission (ITC). Eight organizations participated in this commission to develop twenty-two guidelines for test adaptation that cover (a) the context, (b) test development, (c) administration, and (d) score interpretation (Hambleton, 2005).

Context refers to minimizing the effects of cultural and linguistic factors that could affect the comparison of test scores across cultures, and to ensure that the construct is defined in the same way by the different populations.

Test adaptation refers to the necessary actions that need to be taken in order to ensure that the test is appropriate for the targeted populations. These actions include taking into account the cultural and linguistic differences when considering the item format, content, stimulus, testing techniques and rubrics for the targeted populations. In addition, statistical evidence should be provided to demonstrate the equivalence of the items and test validity.

Test administration includes the different aspects of test implementation that could affect the tests scores validity, such as environment, clear instructions, providing a test manual, and controlling the interaction between test administrator and students.

Finally, *score interpretation* should be supported by evidence of the equivalence of the tests and differences in scores between populations should not be considered as true differences in the construct measured without the appropriate evidence that support test validity.

The guidelines provided by the ITC show that the process of adapting a test should not be reduced to translating from one language to another. Instead proper time and review iterations as well as item piloting should be planned ahead (Solano-Flores & Gustafson, 2013).

Even when students are provided with tests written in their native language, the resulting versions of the tests might differ in difficulty level and/or elicit different interpretations across populations. Wiliam (2008) provides an example of a word used in a PISA item which use is not the same in two languages and it is likely to cue a mathematical response only for one of the two populations. Consequently, the difficulty level and construct measured by the item also changes for both populations.

Solano-Flores, Lara, Sexton, and Navarrete (2001) describe the difficulties teachers encountered translating a set of items from English to Spanish, Chinese and Haitian-Creole. Teachers who were teaching students participating in this study were asked to translate a set of items, in addition to having an external translator who reviewed the translations. During the process, teachers noticed a particular item which included a "gumball machine" in the stem to test students' knowledge of probability. As the discussion developed, teachers discussed that this item might disadvantage Haitian-Creole students who have recently migrated to the U.S. because these machines are

almost nonexistent in Haiti. The results and experiences narrated by the researchers demonstrate the challenges of translating items not only in terms of finding the words that will make the two test versions the same measurement instrument, but also, in finding the contexts and constructs that will make both test versions equivalent.

Even when a test is translated and available for ELLs, it is not easy to identify the language in which students should be tested because not all ELLs have the same proficiency level in their native language or in English (Abedi, 2011). Also, it is not easy to determine if the use of this accommodation will ensure ELLs the opportunity to demonstrate what they know. Solano-Flores , Lara, Sexton and Navarrete (2001) found that when assessing students whose native language is other than English, their responses differ depending on whether they answered an English version of the test, or a version written in their native language. They used a small sample of the NAEP Science and Math items to asses a sample of Spanish, Haitian-Creole and Chinese speakers. Sample items were translated to the three languages, and students answered a version of the items in English and another in their native language. Overall, students didn't perform consistently better when they were tested in their native language, although there was a small group of students who performed consistently better in English or their native language. However for some of the items students performed better in their native language, but for other items, student performance was better in English.

ELLs might benefit from translated tests, if they received content instruction in their native language and are provided with tests written in their native language. But if they received content instruction in English, they might not be familiar with the

vocabulary in their native language, and they might benefit more from an English test (Abedi, 2011). Providing students who received instruction in English a test in their native language might be confusing for them (Abedi et al., 2004; Solano-Flores & Gustafson, 2013).

In summary, research findings show that items might elicit different students' thinking from the one intended by the test developers depending on the way in which test items are written or translated. Consequently, the measure obtained by such items does not correspond to what was originally thought of, posing a threat to scores' validity. In terms of translated tests, there is not still a consensus that helps educators to determine when students should be tested in English or their native language.

Construct Bias

Large-scale assessments: a measure of English proficiency?

Large-scale assessments are often considered tests of language rather than tests of the specific content addressed, and they tell little about what students know (Abedi, 2002; Trumbull & Solano-Flores, 2011). Although it is often thought about ELLs as the main population affected by the language contained in a test, "language plays a critical role in the validity of assessment for every single student" (Trumbull and Solano-Flores, 2011, p. 23). Brislin and Freimanis (2001) also argue that language can be "a source of misunderstanding, even between people who share the same linguistic and cultural background" (p.22). This claim is supported by research reporting how students' interpretations to items affect their performance (Kazemi, 2002; Santel-Parke & Cai,

1997; Verschaffel, De Corte, & Lasure, 1994). The following paragraphs synthesize the main findings on students' English proficiency level impact on their science test scores.

Abedi (2002) studied the effect of K-12 students' language background on achievement tests by comparing the standardized test scores of two student populations – ELL and English native speaker students. He found that ELLs performance was lower than non-ELLs in the three content areas analyzed –reading, math and science; especially in reading, which is a test with higher language demand. For the three content areas analyzed, the gap between both groups was found to be smaller in the lower grades and larger in the higher grades. Moreover analysis of test reliability was found to be lower for ELLs, and the difference between test reliability for non-ELLs and ELLs was higher as the grade level increased.

Luykx et al. (2007) examined how third and fourth grade students' culture and language mediated their science learning using open ended inquiry tasks. After examining 1600 tests, the authors categorized the linguistic influences as follows: (1) orthographical/phonological, and (2) semantic. The former refers to the use of students' native language orthography to write their responses to answers, which according to the researchers can make it difficult for teachers to understand students' responses and be interpreted as lack of knowledge. The later refers to the difficulties in interpreting science terms. Students tend to assign every day meaning to scientific terms, e.g. interpreting the word "states" as the geopolitical states, instead of states of matter (Luykx et al., 2007, p. 909).

The comparison of fifth grade students' responses to 6 multiple-choice items from the Massachusetts state science assessment conducted by Noble et al. (2012) demonstrated that Former Limited English Proficiency (FLEP) students were more likely to misinterpret the question posed in certain items than English native speakers. The difference in performance was noticeable in items with what they called *atypical perspectives* (Noble et al., 2012, p. 792). These are items that introduce situations opposite to those experienced by students in the classroom. For instance, a multiple choice item asked students what would happen to water if heat was taken away from it. According to the researchers, this context is opposite to the most commonly introduced in the classroom, which is leaving water at room temperature or heating it to evaporate (Noble et al., 2012). Data from this study showed that misinterpretation of the items led students to choose an answer choice other than the key to this item during the test, even when they demonstrated knowledge for this item during an interview. Another situation, in which FLEP students were found to underperform in comparison to English native speakers, was when items assessed content that was not taught in school. According to the researchers, English speakers who learned about the content outside school would have an advantage over students who learned the content in a language other than English.

Wolf and Leon (2009), examined the relationship between linguistic characteristics of 542 items from 11 assessments at grades 4,5,7 and 8, and the degree of Differential Item Functioning (DIF) for ELLs. They found variability in linguistic demands across grades with tests being more linguistically demanding in higher grades.

Differential Item Functioning was detected in more items at higher grades, and more items exhibiting DIF were found when the focal group was ELLs with the lowest level of English proficiency than when the focal group was ELLs with high English proficiency level. According to Wolf and Leon (2009) students with low English proficiency are more sensitive to item bias. This might be due not only to their low level of English proficiency, but also because they might be "less attuned to necessary academic culture, and/or classroom factors" (Wolf & Leon, 2009, p. 156).

The research studies described above showed that linguistic factors are likely to be a source of invalidity in ELL testing. Students' performance in science large-scale testing can be affected by their English proficiency level (Wolf & Leon, 2009), and the discourse practices in their native language (Solano-Flores & Nelson-Barber, 2001).

The main implication of the different impact on students' performance depending on their proficiency level, is that making adjustments to tests in terms of the language used and/or providing accommodations to lower the effects of language, might not have the same effects for all students (Abedi, Hofstetter, & Lord, 2004).

In summary, it is likely that ELLs are presented with the challenge of being assessed in a language that it is not theirs. When this happens, they must perform with double effort while answering the test, the first one is trying to understand the item wording, and the second one, using their scientific knowledge to answer the test. Because of this, items might be functioning in two ways: as a measure of student English proficiency, and as a measure of student knowledge. Because measuring students' English proficiency is not the main objective, it can be said that the presence of items that

require from students a high level of English proficiency to endorse them, rather than scientific knowledge is a threat to validity.

PROVIDING VALID ASSESSMENT TO ELLS: ACCOMMODATIONS AND CULTURAL VALIDITY

Recognition of the multiple sources of bias in the assessment of ELLs is a call to improve test validity. Currently, ELLs are provided with accommodations depending on the state education policies. In the case of Texas, Spanish speakers in grades 3-5 are provided with a Spanish version of a test if requested. Nevertheless, Solano-Flores and Nelson-Barber (2001), argue that this measure is limited and does not include a sociocultural perspective considered to be necessary for a deep understanding and inclusion of students' culture into assessment. Research about the effectiveness in accommodations is still in development, and no conclusive results have been reported yet. The following paragraphs present a review of the effects of testing accommodations in students' science test scores and the notion of cultural validity.

Testing accommodations

Testing accommodations for ELLs refer to the support provided to students during a test so they can demonstrate what they know, while leaving non-ELL scores unaltered. Most common accommodations include the provision of a dictionary in students' native language and/or English, reduction of language demand of items, extra time for completing the test and oral directions in students' native language (Butler & Stevens, 2001, p. 413). Providing a translated test is also considered an accommodation. The few

research studies reporting the effects of testing accommodations in students' science scores are described next.

Abedi et al. (2005) conducted a study in which 611 students from fourth and eighth grade -317 students classified as ELL, and 294 as non-ELL, were assigned to four different treatments including the use of any of three accommodations –English dictionary, bilingual dictionary or linguistic modification, or no accommodation. These groups answered released science items from the NAEP. After controlling for English proficiency level, researchers reported that the use of accommodations did increase the performance of ELL students. Dictionaries were found to be more effective among fourth graders, while linguistic modifications were found to be more effective among eighth graders. Overall, test validity was not compromised by the use of accommodations (Abedi, Courtney, Mirocha, Leon, & Goldberg, 2005). A similar result was obtained in a pilot study in which a customized dictionary containing only the words used for the science test was provided to 8th grade students (Abedi, Lord, Kim, & Miyoshi, 2001).

Cultural validity

Attending to the complexity of the interaction between language and culture, Solano-Flores and Nelson-Barber (2001) call for a shift in paradigm in science assessment. They argue that a sociocultural perspective of science assessment is needed to meaningfully address the influence of language and culture in assessment. The inclusion of cultural validity as another dimension to test validity is suggested as an alternative to the cultural and language issues in ELL assessment. Cultural validity is defined as follows:

> The effectiveness with which science assessment addresses the sociocultural influences that shape student thinking and the ways in which students make sense of science items and respond to them. These sociocultural influences include the sets of values, beliefs, experiences, communication patterns, teaching and learning styles, and epistemologies inherent in the students' cultural backgrounds, and the socioeconomic conditions prevailing in their cultural groups. (Solano-Flores & Nelson-Barber, 2001, p. 555)

Requiring that tests be culturally validated would improve the ways in which students' epistemologies, language, cultural views, communication and socialization styles and life context and values are considered for assessment purposes (Solano-Flores & Nelson-Barber, 2001).

The inclusion of cultural validity is, according to Solano-Flores and Nelson-Barber (2001) a way to include ELLs in the construction process of a test. It is not a remedial approach as the use of dictionaries or glossaries; instead, it is a way to ensure that ELLs culture and language are represented in the assessment design process, and that the further use of accommodations is not necessary.

The use of accommodations and the inclusion of cultural validity described in the paragraphs above, are intended to provide ELLs with equitable assessment. However they might also raise issues of validity, comparability, and feasibility. Validity and comparability are compromised with the use of accommodations because non-ELLs scores might also be affected, making comparability of scores no longer possible. In the

case of cultural-validity, the inclusion of language and culture in the test design can yield a version that is not comparable with an English version.

In terms of feasibility, Solano-Flores and Nelson-Barber (2001) point out the rich and deep knowledge of the culture and language needed to design culturally valid assessments, and the long process that constructing these types of tests might represent. Thus designing culturally valid tests for each ELL student population might be cost-inefficient –although this approach is worth trying and could potentially ensure equitable assessment.

Beyond the use of accommodations, researchers call for the participation of ELL experts in the planning, and design of items and assessments to the same extent as teachers and test developers for non-ELL assessment do (Abedi et al., 2004; Kopriva, 2008; Solano-Flores & Trumbull, 2003). The inclusion of ELL experts should change the overall test development process as it would consider the inclusion of ELL populations in the norming process, writing and reviewing items that are culturally sound, and deciding which items should be included in the final version of the test, among other aspects.

Summarizing, it can be said that it is not an easy task to determine which accommodation is more appropriate for ELL students, as is shown by the inconclusive research results in terms of the use of accommodations that have been reported in the literature to date. In this sense, more evidence is needed to better support the decisions of whether certain accommodations control for validity threats, or the ways in which

construct and/or item biases influence students' responses so these biases can be better controlled.

The theoretical framework provided the guide to look for validity evidence of the 2009 fifth grade Science TAKS scores, and also served as the context to situate and interpret the results in this study. This research study aims to add evidence to the body of research discussed above, and to encourage a more systematic study of the ways in which construct and item bias are embedded in items.

Chapter III: Methodology

INTRODUCTION

Findings from the literature reviewed in Chapter II indicate that there are different sources of bias when large-scale assessments are implemented with students from diverse cultural and linguistic backgrounds, making difficult to show what students know. In biased assessments, items might be introducing additional difficulty sources that are not relevant to the construct being measured; making test scores less valid for certain student groups (Camilli & Shepard, 1994). Detecting and eliminating bias when assessing ELLs, is of especial relevance when high-stakes decisions are made based on test scores (Hambleton & Rodgers, 1995). In the U.S. this is considered a national priority given the rapid demographic changes (Turkan & Liu, 2012). This research study examines students' Science scores in the 2009 fifth grade English and Spanish versions of the Texas Assessment of Knowledge and Skills (TAKS) and the equivalence of the English version structure across three student groups from different ethnic and linguistic backgrounds. Specifically this study explores the constructs assessed by the 2009 fifth grade Science TAKS –English version, and identifies possible biased items –items that introduce characteristics resulting in a differential performance between English and Spanish speakers. The fifth grade 2009 Science TAKS was chosen for the following reasons: (a) 2009 is the most recent year of data available for which TAKS items were released allowing for a more detailed item analysis, (b) fifth grade is the first school year in which Texas students are tested in Science using a large-scale assessment, and (c) fifth

grade is the only school year in which ELLs in Texas can be provided with a Spanish version of the Science TAKS.

This study draws from the research examining the possible sources of bias in testing of ELLs presented in Chapter II, and is conducted under Messick's validity framework. Considering validity as an "integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment"(Messick, 1989, p. 13), yields to a research study in which the examination of whether Hispanic students in Texas who might or might not be limited English proficient are being provided with valid assessment implies a continuing process in the look of evidence to assess whether test scores reflect what is expected about this group of students and the decisions that will be made on this basis.

RESEARCH QUESTIONS

The research questions guiding this study are:

a. Are there differences in science scores between fifth grade students from different ethnic and linguistic backgrounds who answered the English or Spanish versions of the 2009 Science Texas Assessment of Knowledge and Skills?

b. Is the probability of endorsing the 2009 fifth grade Science Texas Assessment of Knowledge and Skills items the same for students who answered the English or Spanish version of the test?

c. Are the constructs measured by the 2009 fifth grade science Texas Assessment of Knowledge and Skills equivalent across students from different ethnic and linguistic backgrounds?

It is not the purpose of the study to say whether the Spanish version of the 2009 Science TAKS is a valid instrument to assess ELL students, but rather, to take a first step into examining the possible differences in science scores between cultural and linguistic diverse students, and whether possible sources of construct or item bias exist that might be leading to the potential difference in scores between the four groups identified in this study.

RESEARCH DESIGN

Validity evidence for the 2009 fifth grade Science TAKS for Spanish speakers' scores will be generated in two stages. The first stage is aimed to identify any gaps in the 2009 fifth grade science TAKS scores between Hispanic and Anglo students who answered the Spanish or English version using a one-way ANOVA. It is important to say that the ANOVA is conducted as an exploratory analysis to detect possible differences in performance between the four groups identified for the analysis described in detail in the Participants section, but not to make claims about possible causes of such differences.

The second stage of the design is constituted by two analyses that can reveal any item and/or construct bias of the Science TAKS scores: (1) Mantel-Haenzel statistic, and (b) Confirmatory Factor Analysis with multiple groups (CFAMG). Mantel-Haenzel statistic is commonly used for detecting differences in the conditional probability of endorsing an item for different groups. CFAMG is used to identify possible lack of equivalence in the

constructs assessed by an instrument among different groups (e.g. Brown, 2003; Ercikan & Koh, 2005). A synopsis of both stages in the research design for this study is provided in Figure 2.

	Stage1	Stage 2	
Research Question	a. Are there differences in science scores between fifth grade students from different ethnic and linguistic backgrounds who answered the English or Spanish versions of the 2009 Science Texas Assessment of Knowledge and Skills?	b. Is the probability of endorsing the 2009 Fifth grade Science Texas Assessment of Knowledge and Skills items the same for students who answered the English or Spanish version of the test?	c. Are the constructs measured by the 2009 Science Texas Assessment of Knowledge and Skills equivalent across students from different ethnic and linguistic backgrounds?
Methodology	Analysis of Variance (ANOVA)	Mantel-Haenszel statistic	Confirmatory factor analysis with multiple groups
Purpose	To identify any possible difference in science scores between groups	Determine if the probability of response is the same for the identified groups conditioned on the observed score	Compare the construct equivalence between test versions
Variables	Dependent Variable: -2009 fifth grade Science TAKS score Independent Variable: - Group (based on all possible combinations of: Ethnicity (White, Hispanic), LEP, and Science Test Language.	Students responses to individual items	

Figure 2: Research design

Instrument: 2009 fifth grade Science Texas Assessment of Knowledge and Skills

The Texas Assessment of Knowledge and Skills (TAKS) is the state assessment administered by TEA used to show the degree to which students are learning the state-mandated curriculum known as Texas Essential Knowledge and Skills (TEKS). The Science TAKS given at fifth grade is based on the TEKS from grades 2, 3, 4 and 5 (Texas Education Agency [TEA], 2004).

TAKS is organized using test objectives or "umbrella statements" (TEA, 2004) that group the standards or TEKS in a meaningful way. The four objectives assessed in the fifth grade Science TAKS are: (1) Nature of Science, (2) Life Science, (3) Physical Science, and (4) Earth/Space science. The total number of items assessing these objectives is 40.

In terms of formatting, items are written as multiple-choice with four options. Some of the items are part of a cluster; this is a series of items which make reference to the same initial passage.

The development of the items is a fifteen stage process that starts with the development of the test objectives based on the state-mandate curriculum. After test objectives are revised, prototype items are written and piloted with Texas students, and a blueprint is developed specifying the number of items per objective. The following stages include the revision of items by educators, TEA curriculum and assessment specialists and field-testing with representative samples of Texas students. Finally studies on reliability and validity are conducted (TEA & Pearson, 2010).

The TAKS items for the Spanish version are developed following two processes. Some of the items are transadapted. In this process, the original English text is translated to Spanish considering possible linguistic and cultural differences. The result is a series of culturally rich items that "reflect the terminology and language of the… textbooks and classrooms" (Zucker, Miska, Alaniz & Guzmán, 2005, p. 6). Transadaptation involves several iterations, in which the items are revised by translators, content experts, assessment specialists, bilingual educators and the Texas Education Agency. The other part of the items is originated in Spanish. These items are independently developed and their revision follows the same guidelines used for the revision of items originated in English. As a result of the way in which the Spanish version is developed, the Spanish and English versions of the 2009 Science TAKS only have 11 items in common. There are only 11 Spanish items that are a direct translation of the English version.

Student scores are reported in two ways. The raw score corresponds to the total number of correct items, and the scale score corresponds to a transformation of the raw score into a scale score taking into account the difficulty level of the set of questions in which the score is based.

Data set

Data used for this study were provided by the Texas Education Agency. The data set includes information related to: (1) Administration and student identification, (2) Demographics, (3) Accommodations, and (4) Scores on English, Mathematics, and Science. The variables used for this study are described in Table 1.

	Variable	Description	Code
Demographics	Sex	Student gender	-Male -Female
	Ethnicity	Student ethnicity	-American Indian or Alaskan Native -Asian or Pacific Islander -African American -Hispanic -White-non Hispanic
	Economic Disadvantage	Whether student is eligible for free or reduced price meals	-Not eligible for free or reduced-priced meals -Eligible for free or reduce-priced meals
	Gifted and talented	Student participates in a gifted/talented program	-Yes -No
	Special Education	Student receives special education	-Yes -No
	LEP-Indicator-Code	Whether the student is considered limited English proficient (LEP), this is, if the student is learning to speak in English	-Student currently identified as LEP or has met the criteria for bilingual or ESL program. -Non-LEP student
Science Scores	Science language version	Language in which the Science test is provided	-English -Spanish
	Science test version	The Science test version answered by the student	-TAKS (Accommodated) -TAKS -LAT -TAKS-M -TAKS-Alt
	Science online testing administration	Whether the student answered the test online	-Yes -No
	Science scale score	Student test score	
	Science Item Student Responses	Student responses to individual items	-Correct -Incorrect
	Science lat info	Whether the student was provided with any type of linguistic accommodation including bilingual dictionary, glossary, reading assistance, oral translation or linguistic simplifications	-Yes -No

Table 1: TAKS variables and coding

The coding of two variables was modified for the purpose of this study. Economic disadvantage is coded by TEA (2009a) a four level nominal variable. The first level

refers to students who are eligible for free meals, the second level refers to students who are eligible for reduced price meals, the third level refers to other types of economic disadvantage, and the fourth level refers to students who are not identified as economically disadvantaged. Economic disadvantage was re-coded into two levels. The first level comprises any type of economic disadvantage –aggregating the three levels previously described, and the second level refers to students not identified as economically disadvantaged.

Limited English Proficient Indicator-Code (LEP-Indicator-Code) is also a four level nominal variable (TEA, 2009a). The first level refers to students that are identified as Limited English Proficient (LEP), the second level refers to students that are no longer consider LEP and are in their first year of bilingual or English as a Second Language (ESL) program, the third level refers to students in their second year in a bilingual or ESL program, and the fourth level refers to students that are not LEP. For this study, the LEP-Indicator-Code variable was re-coded into two levels. One that considers LEP and students in their first or second year in a bilingual or ESL program, and the second level refers to non-LEP students. This grouping is based on the claim that bilingual students do not necessarily have the same proficiency level in their two languages (Solano-Flores & Gustafson, 2013), and even if they are participating in bilingual or English as a second language (ESL) programs their English proficiency might not be the same as their native English speaker counterparts.

Participants

The participants were fifth grade students who took the written version of the 2009 Science TAKS. Students were classified in eight groups according to the different possible combinations of the following variables: (a) ethnicity, (b) test language, and (c) student Language English Proficient classification (see Table 2). This initial study focuses only on the written TAKS. Therefore, students who answered other than the written version of the TAKS (e.g. online, TAKS-M, etc.), participated in a gifted and talented program, or received special education, as well as students who received any kind of linguistic accommodation during the test (e.g. linguistic simplifications, oral translation, reading assistance, bilingual dictionary or glossary), were excluded from the analysis to reduce the variables that could potentially contribute to differences in performance between student groups. According to the characteristics considered for student selection, there could be eight different groups (the estimated population for each of the groups according to TEA (2009b) is indicated in Table 2). From the groups indicated in Table 2, I only considered groups 1 to 4 because the population of groups 5 to 8 is negligible, with less than 0.15% of the total population.

Group	Ethnicity	Science test language	LEP students	Estimated population
1*	White	English	No	100,000
2*	Hispanic	English	No	100,000
3*	Hispanic	Spanish	Yes	3,000
4*	Hispanic	English	Yes	40,000
5	Hispanic	Spanish	No	0
6	White	English	Yes	348
7	White	Spanish	Yes	0
8	White	Spanish	No	0

Table 2: Characterization of the groups used for the study
* Groups considered for the study

Sample selection

After identifying the four groups according to the characteristics described in Table 2, a random sample was selected from each of the groups. The sample size for each group was n=1,116, which provided a confidence interval at 95% with a margin error of ± 3%.

DATA ANALYSES

Stage 1: Analyzing differences across groups

To answer the research question: "Are there differences in science scores between fifth grade students from different ethnic and linguistic backgrounds who answered the English or Spanish versions of the 2009 Science Texas Assessment of Knowledge and Skills?", I conducted a one-way ANOVA with students' scale scores on the fifth grade 2009 Science TAKS as the dependent variable is, and Group as the independent variable. The variable Group is a combination of three variables: Ethnicity,

English proficiency and Test language (see Table 2). ANOVA tests the hypothesis that the observed means for the various groups come from the same normal population (Gamst, Meyers, & Guarino, 2008). Thus, ANOVA showed any possible differences between the groups. The effect of socioeconomic status in this study was not of particular interest, however, its correlation with Group was considered to interpret the results attending to previous studies that have indicated that socioeconomic status is a relevant variable when analyzing student achievement differences (Abedi, Leon, & Mirocha, 2003; Carmona et al., 2011), and not taking it into account could confound the results. In addition, the percentage of students in Group 3 that are considered economically disadvantage is noticeable –approximately 93% of the population. A preliminary analysis of students' science scores variation according to socioeconomic status was conducted before the ANOVA to determine the inclusion of socioeconomic status in the analysis.

Stage 2: Collecting validity evidence

Examining differential item functioning

The purpose of the second stage of the study was to collect evidence of the validity of the fifth grade 2009 Science TAKS. The first research question guiding this stage follows: "Is the probability of endorsing the 2009 fifth grade Science Texas Assessment of Knowledge and Skills items the same for students who answered the English or Spanish version of the test?" This research question was addressed by examining Differential Item Function (DIF). "DIF is said to occur whenever the conditional probability, P(θ), of a correct response differs for two groups" (Camilli &

Shepard, 1994, p. 58), when both groups are matched at an ability level. Hence, examining items for DIF, is a way to identify biased items. There are a variety of DIF methods, although for the purpose of this study I used the Mantel-Haenzel chi-square. The use of this method for detecting DIF is very popular (van de Vijver & Leung, 1997), and according to Camilli & Shepard (1994), it has more advantages than IRT methods, as it requires smaller sample sizes, and the results of the analysis do not depend on the IRT model selected, as it assumes that guessing and discrimination is the same for the compared groups. The Mantel-Haenszel statistic "is used for determining whether two variables are independent of one another while conditioning on a third variable" (de Ayala, 2009, p. 327). The third variable is the observed score. This procedures starts with identifying the groups of interest and the construction of a 2×2 contingency table for an item at each score level. This is, for a given item, the number of contingency tables constructed equals the number of score levels. The science TAKS has 39 score levels, thus, to analyze whether a certain item presents DIF, 39 contingency tables should be constructed, and test if the odds of having a score of one is the same for both groups at the 39 score levels. An example of a contingency table for a jth total score on test is presented on Figure 3.

	Score on item		
	1	0	Total
Group A	A_j	B_j	n_{Aj}
Group B	C_j	D_j	n_{Bj}
Total	m_{1j}	m_{0j}	T_j

Figure 3: Example of contingency table

After the contingency tables are constructed, the Mantel-Haenzel statistic is calculated. It is a modified chi-squared statistic with the null hypothesis that the odds for answering 1 are the same for the reference and focal groups.

Based on the claim discussed in chapter 2 that it is likely that large-scale assessments are typically constructed to assess white middle class English speakers (Butler & Stevens, 2001; Laosa, 1977), this study considered the following pair wise comparisons for which the Mantel-Haenzel statistic was calculated: (a) Group1 - Group 2, (b) Group 1- Group 3 and (c) Group 1 - Group 4, (d) Group2 - Group 3, (e) Group 2 - Group 4, and (f) Group 4 - Group 3 (Groups defined in Table 2).

Analyzing construct bias

The third research question guiding the second stage of the research study follows: "Are the constructs measured by the 2009 fifth grade Science Texas Assessment of Knowledge and Skills equivalent across students from different ethnic and linguistic background?" This question reflects the need to examine the construct validity across groups based on the research showing that different constructs might be assessed when

tests designed for middle class, white English students are applied to other ethnic/linguistic populations (Abedi, 2011); and the consideration that a translated test might differ in the constructs being measured (Sireci, Patsula, & Hambleton, 2005; van de Vijver & Leung, 1997). For this purpose, a confirmatory factor analysis with multiple groups (CFAMG) was conducted. This analysis allows determining if the hypothetical structure of the test fits the empirical data across groups, and if language can be considered as another dimension when testing Hispanic and/or ELL students.

CFAMG was selected over exploratory factor analysis (EFA) due to its capability to assess construct equivalence across groups by examining whether the hypothesized factor structures are similar for all groups (Brown, 2006). CFAMG tests for: (a) configural equivalence and (b) measurement equivalence. *Configural equivalence* refers to whether the baseline model –this is the number of factors and item groups follow the same structure across groups (Byrne, 2008).

If the theoretical configuration holds across the four groups for the study, then *measurement equivalence* is tested for. This test consists in estimating the factor loadings for the first group, and constrained equal for the other groups. Then the fit of the model is compared across groups.

In summary, the methodology described was used to identify possible differences across the target student groups, while the Mantel-Haenzel statistic and the CFAMG is conducted to assess item and construct bias.

Chapter IV: Results

This study addresses the three following research questions:

a. Are there differences in science scores between fifth grade students from different ethnic and linguistic backgrounds who answered the English or Spanish versions of the 2009 Science Texas Assessment of Knowledge and Skills?

b. Is the probability of endorsing the 2009 Fifth grade Science Texas Assessment of Knowledge and Skills items the same for students who answered the English or Spanish version of the test?

c. Are the constructs measured by the 2009 Science Texas Assessment of Knowledge and Skills equivalent across students from different ethnic and linguistic backgrounds?

The following pages present the results of the quantitative analyses conducted to answer these research questions, and provide a detailed description of the sample used to address the scope for generalizations of these results.

DESCRIPTIVE STATISTICS

From the population of all 5^{th} grade students who presented the Science TAKS in 2009, a random sample of n=1116 was selected from each of the four student groups considered for the analysis -according to their ethnicity, the language of the TAKS they took and their LEP classification (see Table 2). The randomness of the sample assured that each student within each group had the same chance of being selected for the study, and that it is a representative sample of each population described in Table 2. For

instance, the sample selected represents 1.2 % of the total population of Group 1, 1.5% of the total population of Group 2, 36.2% of the total population of Group 3, and 1.9% of the total population of Group 4. This allows the generalization of the results presented in this chapter to the target population with a confidence interval at 95% and a margin error of 3%, which is considered very robust for educational research standards (U.S. Department of Education & Institute of Education Sciences, 2009). Table 3 shows the distribution of the students with respect to gender and group, indicating that females and males were equally represented in the sample. Table 4 shows the distribution of the students with respect to group and socioeconomic status. It is noticeable the fact that socioeconomic status within each group is homogeneous, especially for groups 3 and 4. More than 86% of Group 1 students are classified as *non-disadvantaged*, 73% of Group 2 students are classified as *disadvantaged*, and more than 98% of Groups 3 and 4 students are classified as *disadvantaged*. Thus, it is worth noting that the group and socioeconomic status variables are confounded, and this was considered in the research design for this study.

Table 5 presents the mean and standard deviation of the Science TAKS scores for each group. Group 2 shows the largest standard deviation, while group 3 is the most homogeneous in test scores.

	Males	Females	Total
Group 1 (White,testlang English,non-LEP)	578	538	1116
Group 2 (Hispanic, testlang English, non-LEP)	527	589	1116
Group 3 (Hispanic, testlang Spanish, LEP)	526	590	1116
Group 4 (Hispanic, testlang English, LEP)	586	530	1116

Table 3: Students' gender by group

	Non-Disadvantaged	Disadvantaged	Total
Group 1 (White,testlang English,non-LEP)	963	153	1116
Group 2 (Hispanic, testlang English, non-LEP)	292	824	1116
Group 3 (Hispanic, testlang Spanish, LEP)	2	1114	1116
Group 4 (Hispanic, testlang English, LEP)	15	1101	1116

Table 4: Students' socioeconomic status by group

	n	Mean	Std. Deviation
Group 1 (White, testlang English, non-LEP)	1116	2383.88	272.00
Group 2 (Hispanic, testlang English, non-LEP)	1116	2261.39	297.87
Group 3 (Hispanic, testlang Spanish, LEP)	1116	2035.77	245.97
Group 4 (Hispanic, testlang English, LEP)	1116	2176.97	260.63

Table 5: 2009 Science TAKS scores by group

EXAMINING DIFFERENCES BETWEEN GROUPS: ANOVA RESULTS FOR RESEARCH QUESTION 1

To determine if there were any differences in Science scores between the four groups considered for the study (see Table 2), a one-way analysis of variance (ANOVA) was conducted using Science TAKS scale scores as a dependant variable. The purpose of this analysis was to look at differences between the identified groups instead of determining the amount of variance accounted by certain variables (ethnicity, gender, socioeconomic status, etc.) from TAKS Science scores' variability. This decision corresponds to the overall purpose of this dissertation to focus on language and culture in the 2009 Science TAKS students' responses. Thus, grouping students according to their ethnicity, English language proficiency and the language in which they took the 2009 Science TAKS was considered to be appropriate for such purpose. Nevertheless, socioeconomic status was identified as a variable that needs to be controlled for, given previous analysis of the literature (Abedi, Leon, & Mirocha, 2003; Carmona et al., 2011). However, the descriptive statistics from Table 4 suggested that including it would not help to explain more score variance existing across groups, than the one already being explained by the grouping variable, as group and socioeconomic status are confounded - as suggested by the percentage of students classified as disadvantaged or non-disadvantage in each group – 86% of Group 1 are non-disadvantaged, 73% of Group 2 is disadvantaged, and 98% of Groups 3 and 4 are disadvantaged. Therefore, it is important to keep in mind that the differences in Science scores might also be generated by diverse factors related to students' socioeconomic status or the resources available at the schools

(Lee & Buxton, 2010). The analysis of the influence of such factors in student achievement differences is beyond the scope of this study.

The ANOVA designed for this study considered Group as the only independent variable, and 2009 Science TAKS scale scores as dependent variable. An initial screening of the data was conducted for model adequacy, eliminating 118 outliers from the sample, which averaged 38 outliers per group. The following assumptions for ANOVA were checked for: (a) independence of errors, (b) normal distribution of errors and (c) equal variances across student groups. The Shapiro-Wilk test was used to check the normal distribution of errors, as it allows detecting departures from normality. Gamst, Meyers and Guarino (2008) suggest using an alpha level for this test of $p < 0.001$ before assuming departure from normality. Shapiro-Wilk test was statistically significant ($p < 0.001$) for each level of the independent variable, indicating that the distribution of errors associated with TAKS scores depart significantly from normality. Equal variance across groups was checked using the Bartlett test. An alpha level of $p < 0.05$ was used for this test before assuming heterogeneity of variance. Bartlett test was statistically significant ($p < 0.001$), suggesting that the data does not meet the assumption of homogeneity of variance. Even when these two assumptions were violated, I proceeded to conduct an ANOVA because of its robustness, and the fact that the same results were obtained using the nonparametric Kruskal-Wallis test.

With the ANOVA, a significant difference in scores was found, $F(3, 4342) = 590.6$, $p<0.001$. Subsequently, pairwise comparisons were conducted. Two procedures for pairwise comparison were used, as homogeneity of variance could not be assumed

according to the results obtained from the Bartlett test. The first procedure assumed homogeneity of variance (Tukey). However, since homogeneity of variance could not be assumed given the results obtained from the Bartlett test, a second analysis was conducted assuming that the groups did not have comparable variances (Tamhane). The significance level for both test was set at $p < 0.05$. Both procedures reported similar results, indicating the existence of significant score means differences ($p<0.05$) between each pair of Groups compared. Differences in scores between the four groups, revealed that Group 1 scored the highest, followed by Group 2, 4 and 3. See Table 6 for the differences in scores between groups.

Groups compared	Difference in Science scale scores means	95% Confidence Interval	p
Group2-Group1	-114.21	[-136.76, -91.65]	< 0.001
Group3-Group1	-357.07	[-379.74, -334.40]	< 0.001
Group4-Group1	-214.69	[-237.33, -192.05]	< 0.001
Group3-Group2	-242.87	[-265.58, -220.15]	< 0.001
Group4-Group2	-100.48	[-123.16, -77.80]	< 0.001
Group4-Group3	142.38	[119.59, 165.18]	< 0.001

Table 6: Differences in Science mean scale scores across groups

The results from Table 6 indicate that there are differences in the TAKS mean scale scores between the four groups considered for the analysis, with the Group 1 (White, testlang English, non-LEP) scoring higher, and Group 3 (Hispanic, testlang Spanish, LEP) scoring the lowest. These results are consistent with the achievement gap between different ethnic and linguistic groups that has been reported (National Center for Education Statistics [NCES], 2011).

ITEM FUNCTIONING: MANTEL-HAENSZEL RESULTS FOR RESEARCH QUESTION 2

In order to study if the probability of endorsing an item for students belonging to each of the four groups is the same, a differential item functioning was performed between the four student groups included in this study. The Mantel-Haenszel (MH) procedure is one of the statistical methods commonly used for detecting uniformly biased items in large-scale testing (Camilli & Shepard, 1994), and allows identifying whether students from different groups have the same probability of endorsing an item. If students from the two groups compared are found to have the same probability of endorsing the item, then the item is said to function similarly for both groups from a psychometric perspective.

Each of the forty items from the 2009 fifth grade Science TAKS was examined for DIF using the MH method. The DIF analysis was conducted using the package *difR* (Magis, Beland and Raiche, 2013) available in R. For this analysis, students were grouped according to their observed score which is dichotomous (0 or 1) for individual items, and the probability of endorsing the item was calculated across groups. Items were flagged using Educational Testing Service (ETS) DIF criteria based on the MH chi-square statistic, according to two categories to classify items: A, B or C, and a + or - sign. The + sign indicates that the odds of endorsing the item are higher for the focal group –item is more difficult for the reference group, while a – sign indicates that the odds of endorsing the item are higher for the reference group. The A, B and C levels depend on the Mantel-Haenszel delta difference (MH D-DIF) statistic and its statistical significance. The MH D-DIF is a transformation of the natural logarithm of the odds ratio of answering

correctly for the reference and focal groups. For instance, if the MH D-DIF statistic has a value of 1 it would mean that the odds of answering correctly are approximately 50% higher for the reference group than for comparable members of the focal group. If the MH D-DIF statistic has a value of 1.5, it would mean that the odds of answering correctly are approximately 90% higher for the reference group than for comparable members of the focal group. This transformation creates a scale used to classify the items according to their DIF level. For A-level items the chi-square statistic is not significant at the 5% level *or* the absolute value of the MH D-DIF is smaller than 1- this is considered a small effect size. C-level items have a chi-square statistic significantly greater than 1 at the 5% level and the absolute value of the MD D-DIF must be greater than 1.5 –this is considered a large effect size. B-level items are considered those that do not fall in the A or C levels and meet two conditions: the chi-square statistic is greater than 3.84, and the absolute value of the MH D-DIF is greater than 1. B-level items require revision but are not considered seriously flawed. From this classification it is important to notice that items flagged at level A, would not be considered biased if the MH statistic is not statistically significant.

I conducted six DIF analyses that can be grouped in three sets (see Table 7) to conduct all possible pairwise comparisons: three DIF analyses had Group 1 as the reference group, two DIF analyses had Group 2 as reference, and one DIF analysis had Group 4 as the reference group (see Table 2 for group descriptions). When comparing any two groups, I chose the group who answered the English version of the TAKS test as the reference group, because this group represents the majority of the student population.

Table 7 presents a summary of the DIF analyses conducted. The last column indicates the variables that are different for the two groups involved in each analysis. For instance, in the DIF analysis including Groups 1 and 2, these groups only differ in ethnicity; they answered the same test version and are classified as non-LEP, thus the variable indicated in the last column is *ethnicity*. DIF analysis in which groups only differ in one variable are of especial interest as long as such variable could be considered as one of the main factors contributing to DIF. DIF analyses in which the groups differ on two or three variables, also speak to the validity and equivalence of the items.

For these analyses, it is important to notice that only 11 items from the 40 in the 2009 fifth grade Spanish Science TAKS are a direct translation of the 2009 English Science TAKS (both versions of the TAKS can be retrieved from TEA website). The other 29 items are not a one to one correspondence with those from the English test version. Thus, the DIF analyses involving group 3 could only be conducted using the common items, since this group answered the 2009 fifth grade Spanish Science TAKS. The number of items used for each DIF analysis is indicated in Figure 4, and it is 40 for groups who answered the English version of the 2009 fifth grade Science TAKS, and 11 for analyses that involve the Group 3, which answered the Spanish version of the TAKS.

DIF Analyses		Variables in which the groups differ
Reference Group	Focal Group	
1	2	Ethnicity
1	3	Ethnicity, Test language, LEP classification
1	4	Ethnicity, LEP classification
2	3	Test language, LEP classification
2	4	LEP classification
4	3	Test language

Table 7: DIF Analyses

Figure 4 presents a synthesis of the number of items used in each analysis, the total number of DIF items identified in the analysis, and the number of items that advantage each population.

	DIF Analyses					
	G1-G2	G1-G3	G1-G4	G2-G3	G2-G4	G4-G3
Total number of items considered for the analysis	40	11	40	11	40	11
Total number of items biased	6	7	24	5	21	24
Number of items that advantage the reference group	4	4	7	4	12	12
Number of items that advantage the focal group	2	3	17	1	9	12

Figure 4: DIF results summary

In the following sections I present the results of the DIF analyses grouped according to the reference group, followed by a section with a discussion of the items that were found to be of interest across the six DIF analyses conducted.

Reference group 1: White, non-LEP students, English TAKS version

DIF analysis Group 1(White, testlang English, non-LEP) and Group 2 (Hispanic, testlang English, non-LEP)

The Mantel-Haenszel analysis conducted between Groups 1 and 2 provided evidence regarding the influence of ethnicity in students' responses to items, because Groups 1 and 2 only differ in ethnicity.

Tables 8 to 10 present, the Mantel-Haenszel chi square statistic, the corresponding ETS classification, the Mantel-Haenszel delta difference (MH D-DIF), item number, and the objective assessed, for each of the items for which the Mantel-Haenszel test was statistically significant, identifying biased items. The MH D-DIF statistic is used to calculate the difference in odds between the focal and the reference group. If the MH D-DIF statistic has a negative sign that means that the odds of endorsing the item are higher for the reference group, and if the MH D-DIF statistic has a positive sign that means that the odds of endorsing the item are higher for the focal group.

ETS Category	MH chi-square statistic	MH D DIF	p-value	Item number(s)	Objective assessed
A-	5.43	-0.74	<0.05	13	4
B+	13.39	1.02	<0.05	18	4
B-	4.73	-1.01	<0.05	12	2
	14.95	-1.00	<0.05	25	4
	11.65	-1.18	<0.05	29	3
C+	10.96	1.66	<0.05	3	3

Table 8: Fifth grade Science TAKS items exhibiting DIF for Groups 1 and 2

The results of the Mantel-Haenszel for Groups 1 and 2 (Table 8) show that from the 40 items in the analysis, 6 items (number 13, 18, 12, 25, 29, and 3) are considered biased. Four of these items (number 12, 13, 25, and 29) advantage Group 1, while only 2 items (number 3 and 18) advantage Group 2. Additionally only one item (number 3) can be considered highly biased and would have received recommendation to be removed from the test according to ETS classification, while the four items flagged with B (number 18, 12, 25 and 29), would have been recommended for revision. The C+ flagged item (number 3), advantages the focal group, which in this case is the Hispanic non-LEP group who answered the English version of the test. The item is presented below.

Item 3

3 A student plays the cymbals in a band. When the cymbals are hit together, a loud sound is produced. The force of the cymbals hitting each other produces sound because —

A metal conducts heat

B energy is absorbed

C metal is magnetic

D air vibrates

Retrieved from http://www.tea.state.tx.us/student.assessment/taks/released-tests/archive/

Figure 5: Item 3 taken from the released 2009 fifth grade Science TAKS

The knowledge intended to be assessed is that a vibrating object can produce sound, classified under the physical science objective (TEA, 2009c).

The small number of biased items (15% of the items) and the balance between those which advantage each group (4 favor the reference group and 2 favor the focal group) suggests that the test as a whole functions psychometrically similarly for groups 1 and 2. That is, from the DIF analysis between Groups 1 and 2 one cannot conclude that the TAKS advantages White over Hispanic non-LEP students when both groups took the 2009 Science English TAKS version.

DIF analysis Group 1 (White, testlang English, non-LEP) and Group 3 (Hispanic, testlang Spanish, LEP)

The Mantel-Haenszel analysis conducted between Groups 1 and 3 provided evidence regarding the psychometric equivalence between items from the English and Spanish versions of the 2009 Science TAKS, and the influence of the interaction of language and ethnicity in students' responses to items. Groups 1 and 3 differ in ethnicity, language, and the test version they answered.

The Mantel-Haenszel analysis for Groups 1 and 3 only considered the 11 items that were a direct translation of the English version of the 2009 Science TAKS. The other 29 items were not used for this analysis, because there is no information regarding the intended correspondence between the English and Spanish versions of the items.

ETS Category	MH chi-square statistic	MH D DIF	p-value	Item number(s)	Objective Assessed
A+	4.81	0.69	<0.05	19	3
B+	5.63	1.06	<0.05	9	3
C+	33.46	1.61	<0.05	21	4
C-	8.32	-2.38	<0.05	1	1
	9.01	-1.64	<0.05	5	3
	23.44	-1.56	<0.05	13	4
	22.60	-1.68	<0.05	29	3

Table 9: Fifth grade Science TAKS items exhibiting DIF for Groups 1 and 3

Of the eleven items analyzed, 63.6% percent were detected with DIF (see Table 9). From the 11 items analyzed, five items (number 1, 5, 13 and 29) are flagged C-, indicating that the advantage group is the White, non-LEP students who answered the English version of the test (Group 1). Only one item (number 21) is flagged C+, indicating that odds of

answering this question are greater for Hispanic, LEP students who answered the Spanish version of the TAKS (Group 3). This item intends to assess students' knowledge of earth materials including rock, soil, water and gasses, to classify them in renewable, non-renewable, and inexhaustible resources (TEA, 2009c). A more detailed discussion of this item and items number 5 and 13 is presented in the *Item analysis* section. Item 29 is illustrated in Figure 6.

<table>
<tr><td colspan="2">Item 29</td></tr>
<tr><td>English Version</td><td>Spanish Version</td></tr>
<tr><td>29 Which two properties of a crayon will stay about the same after the crayon is melted?

A Shape and physical state
B Temperature and hardness
C Color and mass
D Thickness and texture</td><td>27 ¿Cuáles son las dos propiedades de un crayón que seguirán casi iguales después de que se derrita?

A Forma y estado físico
B Temperatura y dureza
C Color y masa
D Grosor y textura</td></tr>
<tr><td colspan="2">Retrieved from http://www.tea.state.tx.us/student.assessment/taks/released-tests/archive/</td></tr>
</table>

Figure 6: Item 29 taken from the released 2009 fifth grade Science TAKS

Item 29 intends to assess students' knowledge of physical properties of matter. The context used is about what would happen to a crayon after is melted. The use of crayons is very common in the elementary years of school in the U.S., giving students the opportunity to get familiar with what happens to crayons in a sunny day, their texture, color, etc. However, the use of crayons might not be common for students who are from other countries. For instance, in Mexico City –one of the cities where some LEP students

come from, crayons are not commonly used in elementary school; their use is limited to pre-K or Kindergarten. Also, in low SES Mexican schools, students might not have access to crayons. Consequently, U.S. students who use crayons on a daily basis and are familiar with their physical properties and have observed how such properties change or not in different situations, are advantaged by the context of this item. Additionally, the Spanish word used for crayon, "crayón" is not the common word in Spanish to name crayon. The word that is most used in Spanish for crayon is "crayola". Other aspect that might add difficulty for Group 3 students (Hispanic, testlang Spanish, LEP) is the vocabulary used in the answer choices. Words such as "physical state" ["estado físico"], "hardness" ["dureza"], and "texture" ["textura"], are likely to be learned at school. LEP students, who answered the Spanish TAKS version and didn't have the opportunity of learning this vocabulary at school, might not be able to answer this question.

It is worth noting that three of the four objectives or "umbrella statements" (TEA, 2004) assessed by the TAKS are represented by the seven items in which DIF was found. 57% of the biased items assess objective 3 –understanding of the Physical Sciences.

From the DIF analysis between Groups 1 and 3 one can conclude that the majority (80%) of the items flagged at level C advantage White, non-LEP students who answered the English version of the test. Even when the number of items that are common to the English and Spanish version of the 2009 Science is small, the presence of 6 highly biased items (number 1, 5, 13, 21 and 29) is evidence of the lack of psychometric equivalence of the TAKS for both populations.

DIF analysis Group 1 (White, testlang English, non-LEP) and Group 4 (Hispanic, testlang English, LEP)

The Mantel-Haenszel analysis conducted between Groups 1 and 4 provided evidence regarding the influence of the interaction of language and ethnicity in students' responses to 2009 Science TAKS items. Groups 1 and 4 differ in ethnicity and their English proficiency.

Contrary to the trend of biased items advantaging mostly Group 1, the DIF analysis involving Groups 1 and 4 (see Table 10), show that there are more items that advantage Group 4 (Hispanic, testlang English, LEP).

From the 40 items considered for the analysis, 10 items are identified as highly biased; meaning that, according to ETS DIF criteria, these items should have received recommendation to be removed from the test. Additionally 14 items were flagged at levels A or B. Consequently, nearly half of the items of this test were classified as biased following the Mantel-Haenszel procedure. Four of the 8 items flagged as C+ assess objective 3 –understanding of the Life Sciences, while both items flagged as C-, which advantage Group 1, assess objective 1 –understanding of the Nature of Science.

ETS Category	MH chi-square statistic	MH D DIF	p-value	Item number(s)	Objective Assessed
A+	12.48	0.97	<0.05	21	4
	6.46	0.73	<0.05	22	1
	5.44	0.97	<0.05	35	1
A-	8.65	-0.92	<0.05	13	4
	5.89	-0.99	<0.05	33	2
B+	4.67	1.08	<0.05	10	1
	15.79	1.29	<0.05	16	2
	20.15	1.25	<0.05	19	3
	16.11	1.28	<0.05	26	4
	24.20	1.44	<0.05	28	4
	7.58	1.19	<0.05	34	3
B-	4.31	-1.29	<0.05	8	1
	7.64	-1.08	<0.05	14	3
	5.55	-1.19	<0.05	32	1
C+	25.48	2.61	<0.05	3	3
	19.34	1.95	<0.05	7	2
	18.71	1.96	<0.05	9	3
	33.37	2.04	<0.05	11	3
	86.04	2.79	<0.05	18	4
	29.92	1.74	<0.05	23	4
	22.43	2.10	<0.05	27	2
	10.24	0.67	<0.05	38	3
C-	10.51	-2.23	<0.05	4	1
	11.49	-1.67	<0.05	6	1

Table 10: Fifth grade Science TAKS items exhibiting DIF for Groups 1 and 4

The fact that the Mantel-Haenszel procedure indicated a large amount of items advantaging Group 4 seems to oppose the difference in the mean of Science test scores between groups 1 and 4 summarized in Table 6; and the literature reporting the achievement gap between Hispanic and White students (U.S. Department of Education, IES, & National Center for Education Evaluation and Regional Assistance, 2009), as well

as the influence of language and culture in student achievement (Solano-Flores & Nelson-Barber, 2001; Solano-Flores & Trumbull, 2003). This situation opened a direction for a finer-grained item analysis of their Item Characteristic Curves (ICC). When looking at the item characteristic curves of some of the items flagged with C+ (Figure 7), it is noticed that the advantage for Group 4 students is not uniform, which helps better understand this apparent contradiction, as is further described using items 18, 11 and 23. These items are used for the finer-grained analysis because they have the higher MH chi-square statistic, and they assess two of the objectives (Physical and Earth/Space Sciences) for which more biased items have been detected in the analysis with Group 1 as the reference.

The ICC of an item shows how "changes in trait level relate to the changes in the probability of a specified response" (Embretson & Reise, 2000, p. 46). According to this description, the ICCs of the items illustrated in Figures 7 and 8, the x-axis represents the trait or ability level and it is measured in the range from -4 to 4, with 4 indicating a high ability level. The y-axis represents the probability of endorsing the item, measured from 0 to 1, where values near 1 indicate a high probability of endorsing the item. Figure 7 shows Groups 1 and 4 ICCs for C+ items 18, 11 and 23.

The continuous line in the ICCs illustrated in Figure 7 represents the ICC for Group 1, and the dotted line represents the ICC for Group 4. Consider item 18 illustrated in Figure 7, the advantage is mainly for Group 4 (Hispanic, testlang English, LEP) students with low achievement levels. In the ICCs graph of item 18, the dotted line lies above the continuous line in the theta interval from -4 to 0 approximately, indicating that the

probability of endorsing the item is higher for students in Group 4. However, both ICCs for item 18, intersect at a value of theta near zero, and in the following interval (from 0 to 4) the advantaged group changes to Group 1 (White, testlang English, non-LEP), as the continuous line lies above the dotted line. The advantage in this interval for Group 4 is small as measured by the separation of the ICCs in the interval from 0 to 4. This means that this item does *not* advantage *all* Group 4 students, and that the advantage increases as the ability level decreases in the interval (-4, 0). The way in which the ICCs interact for items 11 and 23 is very similar, advantaging Group 4 at low ability levels, and Group 1 at higher ability levels.

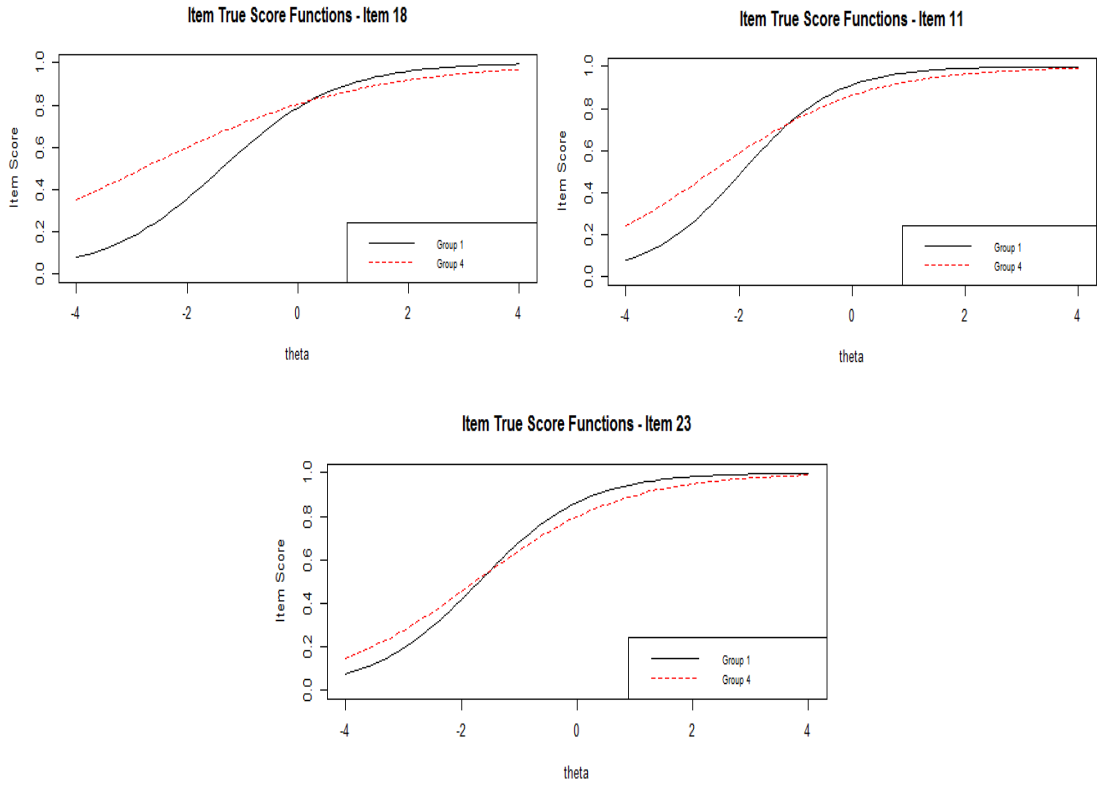


Figure 7: Examples of C+ Item Characteristic Curves for Groups 1 and 4

In summary, the ICCs of the set of items illustrated in Figure 7 show that even when overall items are considered to advantage Group 4 (Hispanic, testlang English, LEP) using the Mantel-Haenszel procedure, when looking at the different ability levels Group 4 is the advantaged group for students with low ability levels by a large margin, while Group 1 (White, testlang English, non-LEP) is the advantaged group for students with high ability levels with a much smaller margin difference.

In contrast, items that advantage Group 1 (White, testlang English, non-LEP) students do so for students at all ability levels, as it is shown in figure 8. Consider the ICCs for item 4 illustrated in Figure 8; at the lower end of ability, Group 4 (Hispanic, testlang English, LEP) students have a probability of approximately 0.1 of endorsing the item, while Group 1 students with the same level of ability have a probability of approximately 0.8 of endorsing the same item. The difference in the probability of endorsing this item between both groups represents an important advantage for Group 1 students. In addition, Group 1 students from all ability levels have almost the same probability of endorsing item 4. When all students have the same probability of endorsing the item, the item does not differentiate between high and low ability students and so, from a psychometric perspective, it is not a very useful item for a test, which one of its objectives is to be able to provide relative comparisons of students' ability levels. Thus, in addition to the differences in ICCs between Groups 1 and 4, psychometrically speaking, item 4 is not a good item for Group 1.

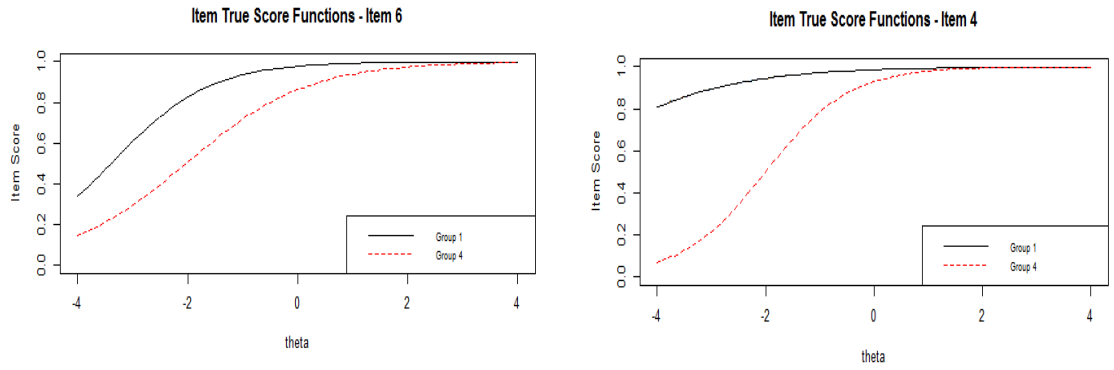


Figure 8: Examples of C- Item Characteristic Curves for Groups 1 and 4

Summary of DIF Results using Group 1 (White, testlang English, non-LEP) as the Reference group

The DIF analyses conducted considering Group 1 as reference showed that the 2009 Science TAKS functions psychometrically similar for Groups 1 (White, testlang English, non-LEP) and 2 (Hispanic, testlang English, non-LEP), suggesting that ethnicity does not influence importantly in the functioning of the items for this two groups. This can be said because the only difference between Groups 1 and 2 is ethnicity. Nevertheless, the DIF between Groups 1 and 2 revealed biased items that would have been recommended for revision according to ETS DIF criteria. In contrast, the DIF analysis between groups 1(White, testlang English, non-LEP) and 4 (Hispanic, testlang English, LEP) showing that almost half of the 2009 Science TAKS items are biased indicates that the items function psychometrically different for Groups 1 and 4, thus it can be said that the interaction of ethnicity and English proficiency influences students' response to items. From the eleven items used in the DIF analysis between Groups 1 and 3, it is noticed that more than half of these items are biased, showing a poor psychometric

equivalence between the English and Spanish items. This lack of equivalence indicates that the interaction of students' culture, native language and test translation affect students' responses to items.

The DIF analysis between Groups 1(White, testlang English, non-LEP) and 4 (Hispanic, testlang English, LEP) revealed three times more biased items than the DIF analysis between Groups 1(White, testlang English, non-LEP) and 2 (Hispanic, testlang English, non-LEP). The fact that Groups 2 and 4 only differ in English proficiency level, adds evidence to support the way in which students' English proficiency level can increase the difficulty of large-scale assessments, making more difficult for students to demonstrate what they know, as it has been previously reported in the literature (e.g. Solano-Flores & Li, 2009, Solano-Flores & Trumbull, 2008).

The three DIF analysis that have Group 1 as the reference show that from the variables considered for grouping students in this study (ethnicity, test version, LEP), the interaction of ethnicity and language is the one that most influence students' responses, resulting in a large number of biased items in the 2009 fifth grade Science TAKS (nearly half the items in the DIF analysis for Groups 1 and 4).

In the three analyses of this set, the majority of items that were classified as highly biased (ETS classification C), whether they advantaged the reference or the focal group, intended to assess students' knowledge of Physical Sciences (e.g. knowing that a vibrating object produces sound). One of the reasons why Physical Sciences might be generating so much variability in item functioning could be that this type of content involves phenomena that are mainly taught in the classroom, such as magnetism,

electricity, sound, etc. Thus, students with limited English proficiency that are exposed to this content in the classroom, have the opportunity to get familiar with the scientific language, increasing their opportunity to perform better in these kind of items than in items that use language that is learned outside the classroom (Moskovich, 2013).

Reference group 2: Hispanic, non-LEP students, English TAKS version

DIF analysis Group 2 (Hispanic, testlang English, non-LEP) and Group 3 (Hispanic, testlang Spanish, LEP)

The Mantel-Haenszel analysis conducted between Groups 2 and 3 will help gathering evidence regarding the influence of the interaction of test language and English proficiency in students' responses to 2009 fifth grade Science TAKS items. Groups 2 and 3 differ in their English proficiency and the version of the TAKS.

Table 11 presents the items that were flagged as biased in the DIF analysis for groups 2 and 3. In this case, only 11 items from the 40 in the 2009 Science TAKS were considered, because only 11 items are a direct translation from the English TAKS version, and thus, in which a direct correspondence was possible between the English and Spanish test versions. From the 11 items, 5 were detected with DIF (number 1, 5, 13, 21, and 29), but only two of them (items 1 and 5) were classified as highly biased advantaging Group 2 (Hispanic, testlang English, non-LEP).

ETS Category	MH chi-square statistic	MH D DIF	p-value	Item number(s)	Objective Assessed
A-	11.52	-0.95	<0.05	29	3
B+	15.44	1.06	<0.05	21	2
B-	24.91	-1.33	<0.05	13	1
C-	18.21	-2.40	<0.05	1	2
	18.06	-1.73	<0.05	5	1

Table 11: Fifth grade Science TAKS items exhibiting DIF for Groups 2 and 3

The five items detected with DIF were also flagged in the DIF analysis between Groups 1 (White, testlang English, non-LEP) and 3 (Hispanic, testlang Spanish, LEP), although for three of them (number 13,21 and 29) the ETS classification is different (see Figure 9). The change in the ETS classification of this items, suggest that ethnicity is a factor that affects students' responses to items 13, 21 and 29, resulting in a higher DIF effect size for these items.

		ETS DIF classification	
Item number	Objective assessed	DIF Groups 1 and 3	DIF Groups 2 and 3
1	1	C-	C-
5	3	C-	C-
13	4	C-	B-
21	4	C+	B+
29	3	C-	A-

Figure 9: Items detected with DIF for Groups 1 and 3 and Groups 2 and 3 analyses

The objectives assessed by the two items flagged C- (number 1 and 5), are related to students knowledge of Nature of Science –objective 1 and Physical Science –objective 3. Both items advantage Hispanic, non-LEP students who answered the English version of

the test, suggesting that the translation of the test might add difficulty to these items for the Hispanic, LEP students who answered the test in Spanish.

DIF analysis Group 2 (Hispanic, testlang English, non-LEP) and Group 4 (Hispanic, testlangEnglish, LEP)

The Mantel-Haenszel analysis conducted between Groups 2 and 4 will help gathering evidence regarding the influence of language in students' responses to 2009 Science TAKS items. The groups considered for this analysis, are the same ethnicity and answered the same test language, though Group 4 is comprised by LEP students.

From the 40 items of the 2009 fifth grade Science TAKS used for the DIF analysis, 21 items were classified as biased (see Table 12). Only 3 items (number 4, 6 and 32) were classified as highly biased (flagged C-), and the three of them advantage Group 2 (Hispanic, testlang English, non-LEP). For the three items flagged at level C-, it could be said that the language adds difficulty to these items.

ETS Category	MH chi-square statistic	MH D DIF	p-value	Item number(s)	Objective Assessed
A+	5.60	0.87	<0.05	9	3
	4.69	0.58	<0.05	16	2
	8.16	0.69	<0.05	19	3
	5.06	0.61	<0.05	26	4
	14.47	0.97	<0.05	28	4
A-	10.56	-0.83	<0.05	13	4
	4.03	-0.68	<0.05	20	2
	4.70	-0.68	<0.05	31	4
B+	11.63	1.19	<0.05	7	2
	11.49	1.03	<0.05	11	3
	25.37	1.32	<0.05	18	4
	14.53	1.01	<0.05	23	4
B-	9.64	-1.41	<0.05	2	2
	6.73	-1.11	<0.05	5	3
	8.02	-1.26	<0.05	8	1
	11.18	-1.01	<0.05	14	3
	20.07	-1.43	<0.05	33	2
	6.29	-1.07	<0.05	37	1
C-	28.50	-2.29	<0.05	4	1
	33.96	-2.19	<0.05	6	1
	30.13	-2.22	<0.05	32	1

Table 12: Fifth grade Science TAKS items exhibiting DIF for Groups 2 and 4

The content assessed by the three items flagged at level C- is students' knowledge of the nature of science. Items assessing this content use language that is used in contexts outside the classroom, making it more difficult for students who don't speak English at home to learn it. Item 4 is illustrated in Figure 10.

Item 4

4 Which tool below best models the way the dog's jaws work?

F 

H 

G 

J 

Retrieved from http://www.tea.state.tx.us/student.assessment/taks/released-tests/archive/

Figure 10: Item 4 taken from the released 2009 fifth grade Science TAKS

The ability of students' answering this item, might depend on their knowledge of how each of the tools represented in the answer choices work, as well as whether they are familiar with the vocabulary used in the stem, such as "jaws" and "tool". These two words are relevant to understanding the question, and are not likely to be taught at school. Consequently a student who does not speak English outside the school might not know these words, making it more difficult to make sense of the question.

The results show that there are more items that advantage non-LEP students; suggesting that language is a factor that influences students' ability to respond to items, as has been found in other research studies (e.g. Martiniello & Wolf, 2012; Noble et al., 2011; Solano-Flores, Lara, Sexton, & Navarrete, 2001).

The DIF analyses considering Group 2 as the reference group, indicate that even when students have the same ethnicity classification by TEA, the Hispanic group is highly heterogeneous, as different students might differ in English proficiency level, and places of origin, which might yield to differences within the Spanish language spoken by students. The fact that most of the biased items for the two analyses of this set, advantage Group 2 (Hispanic, testlang English, non-LEP) speaks not only to the lack of validity of the individual items for LEP students, but also to the relevance of the language in Science testing. The DIF analysis between Groups 2 and 4 indicates that English proficiency is a factor that increases item difficulty, shown by the large number of biased items advantaging non-LEP students over LEP students.

Reference group 4: Hispanic, LEP students, English TAKS version

DIF analysis Group 4 (Hispanic, testlang English, LEP) and Group 3 (Hispanic, testlang English, LEP)

The Mantel-Haenszel analysis conducted between Groups 3 and 4 provided evidence regarding the influence of test language in students' responses to 2009 Science TAKS items. The groups considered for this analysis, are the same ethnicity and English proficiency classification by TEA, though they answered the 2009 Science TAKS in different languages.

Eleven items that are a direct translation from the English version were considered for this analysis. Group 3 answered the Spanish version while Group 4 answered the English version. Three items were flagged with DIF (number 15, 21 and 29), two of them advantaged students who answered the Spanish version of the 2009 Science TAKS (see Table 13). Only one item (number 21) was classified as highly biased according to ETS classification, advantaging students who answered the Spanish version of the test over students who answered the English version of the test. A more detailed analysis of item 21, intended to assess students' knowledge of Earth/Space Science is presented in the following section.

ETS Category	MH chi-square statistic	MH D DIF	p-value	Item number(s)	Objective Assessed
A+	9.09	0.72	<0.05	15	4
A-	5.95	-0.66	<0.05	29	3
C+	36.43	1.58	<0.05	21	4

Table 13: Fifth grade Science TAKS items exhibiting DIF for Groups 3 and 4

Because the number of items that are common to the two language versions of the TAKS represents less than 30% of the total items of the test, it cannot be said whether the Spanish version is an effective accommodation, and whether it is psychometrically equivalent to the English version.

Item analysis

Figure 11 summarizes the items' ETS classification for each DIF analysis. Items common to the English and Spanish versions of the TAKS are highlighted in Figure 11. This set of items is of great interest for this study, because being common to both tests allows the comparison of their functioning through the six DIF analyses conducted.

Figure 11 shows that the number of DIF items varied across analyses and the variability in the psychometric properties of the items and the test across groups. The fact that the ETS classification of some items was not always the same across analyses supports the result that the 2009 Science TAKS items do not function consistently across groups. The variability of item functioning across groups indicates the influence of students' ethnicity, English proficiency and test language in students' responses to items. In terms of ethnicity, the DIF analysis revealed a small number of items between Groups 1 and 2 which only differ in ethnicity; suggesting that the 2009 fifth grade Science TAKS function psychometrically similar between these two Groups. The largest number of biased items were detected in DIF analysis between LEP and non-LEP students: 24 DIF items were detected in the DIF analysis between Groups 1 (White, testlang English, non-LEP) and 4 (Hispanic, testlang English, LEP), while 21 DIF items were detected in the DIF analysis between Groups 2 (Hispanic, testlang English, non-LEP) and 4 (Hispanic, testlang English, LEP). Regarding test language, the DIF analyses were limited to the small number of items (11) that were common to the English and Spanish versions of the test. The DIF analysis between Groups who answered different test versions revealed that small number of biased items were classified with DIF (3 items) when the groups differed in the test version (Groups 4 and 3), and similar number of items (5 items) were classified with DIF when the groups differed in English proficiency (Groups 2 and 3), and English proficiency and ethnicity (Groups 1 and 3). The variation in the number of items in the DIF analyses involving Group 3, provide evidence of the difficulty added by test translation.

		ETS DIF classification					
Item number	Objective assessed	DIF Groups 1 and 2	DIF Groups 1 and 3	DIF Groups 1 and 4	DIF Groups 2 and 3	DIF Groups 2 and 4	DIF Groups 4 and 3
1	1		C-		C-		
2	2					B-	
3	3	C+		C+			
4	1			C-		C-	
5	3		C-		C-	B-	
6	1			C-		C-	
7	2			C+		B+	
8	1			B-		B-	
9	3		B+	C+		A+	
10	1			B+			
11	3			C+		B+	
12	2	B-					
13	4	A-	C-	A-	B-	A-	
14	3			B-		B-	
15	4						A+
16	2			B+		A+	
18	4	B+		C+		B+	
19	3		A+	B+		A+	
20	2					A-	
21	4		C+	A+	B+		C+
22	1			A+			
23	4			C+		B+	
25	4	B-					
26	4			B+		A+	
27	2			C+			
28	4			B+		A+	
29	3	B-	C-		A-		A-
31	4					A-	
32	1			B-		C-	
33	2			A-		B-	
34	3			B+			
35	1			A+			
37	1					B-	
38	3			C+			

Figure 11: ETS classification of the items flagged with DIF across the six Analyses

In Figure 11, items flagged with a minus sign favor the reference group. The highlighted rows indicate the items that are common to the English and Spanish versions of the 2009 fifth grade Science TAKS.

The following section presents the analysis of some DIF items that consistently advantaged non-LEP students, or LEP students. The items considered for analysis in this section were those common to both the English and Spanish version of the 2009 fifth grade Science TAKS, and that were considered highly biased in at least one of the analyses.

Items that advantage non-LEP students

In three DIF analysis (G1-G3, G2-G3, G2-G4), item 5 (see Figure 12) was found to advantage the non-LEP group in each case, and the advantage was greater when Group 3 (Hispanic, testlang Spanish, LEP) was the focal group. Based on this result, it can be said that Spanish translation adds difficulty to this item.

Item 5	
English Version	Spanish Version
5 Ice cream in a bowl changed from solid to liquid in a few minutes. Which of the following most likely caused this change? **A** Bacteria grew in the ice cream. **B** Heat was added to the ice cream. **C** Water evaporated from the ice cream. **D** Frozen berries were sprinkled on the ice cream.	**5** El helado que había en un plato cambió de sólido a líquido en unos pocos minutos. ¿Cuál es la causa más probable de este cambio? **A** Unas bacterias crecieron en el helado. **B** Se calentó el helado. **C** El agua se evaporó del helado. **D** Le pusieron fresas congeladas al helado.
Retrieved from http://www.tea.state.tx.us/student.assessment/taks/released-tests/archive/	

Figure 12: Item 5 taken from the released 2009 fifth grade Science TAKS

Item 5 (Figure 12) asks students to identify why ice cream in a bowl melts. The English version answer is "Heat was added to the ice cream", while the Spanish version correct response is "Se calentó el helado" ["The ice cream was heated"]. Although both options are written in passive voice –which is not recommended for test design, both answers differ in their grammatical structure. While the English version of the item suggests a change in the ice cream through the action of adding heat, the Spanish version suggests that the ice cream was heated, but neither a direct action nor a subject can be identified as the catalyst for adding heat to the ice cream.

The wording "adding" or "take away heat", is similar to the one analyzed in a study by Noble, et al. (2012). In this study, it was found that LEP and students in free or reduced lunch might find it unusual to be talking of phenomena involving melting or cooling substances using the terms "adding" or "taking away" heat. Student' interviews conducted by Noble, et al. (2012) show that students use terms like melting, cooling, turn

into a gas, freezing, etc. to describe changes in state of matter, rather than using words like add or take away heat.

In addition to the terminology, the complexity of the distracters might be adding difficulty to item 5 as they introduce phenomena that might not be familiar or difficult for students to analyze. For instance, the word "bacteria" -introduced in the answer choice A, might not be familiar to fifth grade students.

Item 13 (see Figure 13) also was found to consistently advantage a non-LEP Group in five DIF analyses (G1-G2, G1-G3, G1-G4, G2-G3, G2-G4). The non-LEP groups advantaged in these analyses have different ethnicity classification; Group 1 comprises White students, while Group 2 comprises Hispanic students. The item asks students about the way in which water can be replaced in a lake.

<table>
<tr><th colspan="2">Item 13</th></tr>
<tr><th>English Version</th><th>Spanish Version</th></tr>
<tr><td>13 After lake water flows through a dam, which processes help replace the water in the lake?

A Rainfall and runoff
B Erosion and weathering
C Refraction and reflection
D Separating and evaporating</td><td>13 Después de que el agua de un lago fluye a través de una presa, ¿qué procesos ayudan a reemplazar el agua en el lago?

A Lluvia y escurrimiento de agua
B Erosión y degradación ambiental
C Refracción y reflexión
D Separación y evaporación</td></tr>
<tr><td colspan="2">Retrieved from http://www.tea.state.tx.us/student.assessment/taks/released-tests/archive/</td></tr>
</table>

Figure 13: Item 13 taken from the released 2009 fifth grade Science TAKS

Item 13 exhibited DIF in five comparison analyses. In comparisons between groups who answered the same version of the test (G1-G2, G1-G4, G2-G4) it was classified at A-level; suggesting that this items works psychometrically similarly for students with different ethnicity and English proficiency. However, only in the analysis between Groups 1 (White, testlang English, non-LEP) and 3 (Hispanic, testlang Spanish, LEP) was flagged with C-, suggesting that not only does test language influence students' responses but also English proficiency and ethnicity.

Noticeable is the fact that this item includes terminology that is primarily taught in school, which has been found to reduce item bias due to the inclusion of vocabulary that is not familiar to students (Martiniello, 2013). However, in the case of the translated version, identifying whether this item represents an effective accommodation to Hispanic LEP students will depend on their Spanish proficiency, which students might not develop in school. This result is consistent to what was reported by Solano-Flores, et al. (2001) regarding ELLs not performing necessarily better when tested in Spanish.

In addition, item 13 presents a context that might not be familiar to all students. In the state of Texas, droughts and sources of water are a common topic in daily news, thus, students who have lived in Texas for a long time are very familiar with this topic, which is not necessarily the case for students that are new to the state.

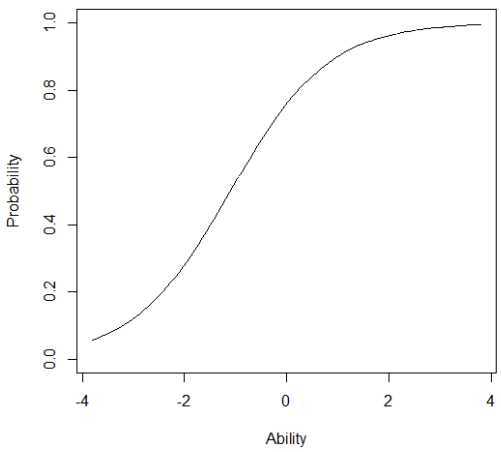
Items that advantage LEP students

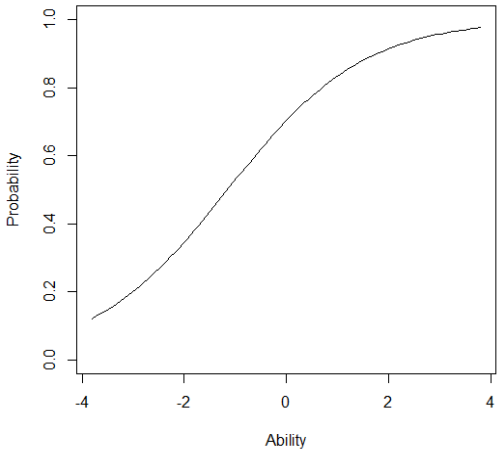
Item 21 (see Figure 14) was detected with DIF in four analyses, in which the advantage Groups were 3 (Hispanic, testlang Spanish, LEP) and 4 (Hispanic, testlang English, LEP). This item is of special interest because in the two languages the stem is

the same for both items, but the choices are different in the English and Spanish test versions.

<table>
<tr><th colspan="2">Item 21</th></tr>
<tr><th>English Version</th><th>Spanish Version</th></tr>
<tr><td>21 Which of the following is best classified as a nonrenewable resource?

A Grass
B Aluminum
C Sunlight
D Oxygen</td><td>17 ¿Cuál de los siguientes recursos se clasifica como un recurso no renovable?

A Carbón mineral
B Árboles tropicales
C Nitrógeno
D Viento</td></tr>
<tr><td> 

Item 21 Characteristic Curve –English Version</td><td> 

Item21 Characteristic Curve – Spanish Version</td></tr>
<tr><td colspan="2">Retrieved from http://www.tea.state.tx.us/student.assessment/taks/released-tests/archive/</td></tr>
</table>

Figure 14: Item 21 taken from the released 2009 Science TAKS

Even when item 21 could be considered to assess the same content for Groups 3 and 4, having different options might change the difficulty level of the question. In fact, after adjusting the 1-PL model to both sets of items separately, I found that the difficulty level is slightly higher for Group 1 (b= -1.102) than it is for Group 3 (b= -1.153), as can also be seen in the lower panel of figure 14. The 1-PL model estimates the probability of an examinee of answering an item correctly given his or her ability level (theta) and the difficulty of the item (b). For this model the difficulty level of the item corresponds to the point in the ability scale where the probability of endorsing the item is 0.5.

The characteristic curves of item 21 show that students at low ability levels have lower probabilities of endorsing the English version of the item than same ability level students who answered the Spanish version. Because of the differences between the item choices of the two items, and the differences in difficulty level, it is not possible to say whether one group of students will know the content better than the other group. However, the answer choices used in the Spanish version: "A. mineral carbon", "B. tropical trees", "C. nitrogen" and "D. wind" might be more familiar to students than the English version choices, as they represent common examples of renewable or non-renewable resources. Aluminum might not be a common example of non-renewable resources. More research is needed to assess the impact of the different choices in students' responses, and whether such impact varies according to students' language and/or culture.

ASSESSING TAKS STRUCTURE: CONFIRMATORY FACTOR ANALYSIS RESULTS FOR RESEARCH QUESTION 3

In order to assess the equivalence between the constructs assessed by the 2009 Science TAKS across the groups considered for this study (see Table 2), a Confirmatory Factor Analysis with Multiple Groups (MGCFA) was conducted. MGCFA was aimed to test if the hypothesized model provided by TEA (Figure 15) fit the empirical data for Groups 1 (n= 1077), 2 (n=1068) and 4 (n=1053), and determine if the items psychometrically assess the same constructs for the three populations. This constitutes a holistic analysis of the test, as it allows analyzing the test structure psychometrically, identifying the construct (s) assessed, as well as how each item loads on each construct. Given that the English and Spanish versions of the 2009 Science TAKS have only 11 questions in common, I considered that it was not appropriate to compare the structure of the English and Spanish versions of the 2009 Science TAKS. Thus, I only assessed whether the structure of the English version of the TAKS holds for Groups 1, 2 and 4 which differ in ethnicity and LEP classification. I used MPlus for this analysis, as it is the only program in which the estimator that has been reported to work best with categorical data is available (Byrne, 2012).

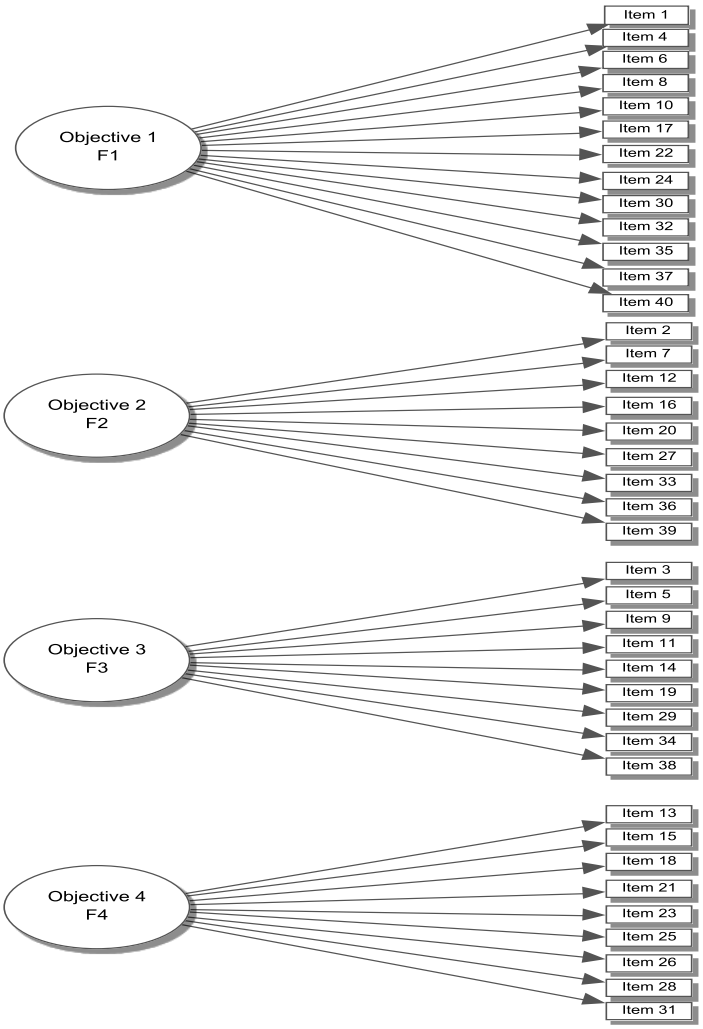


Figure 15: Hypothesized fifth grade 2009 TAKS Science structure from TEA's blueprint

Establishing baseline models

Baseline model for Group 1(Hispanic, testlang English, non-LEP)

The first step to conduct the MGCFA is to establish a baseline model for each Group. Goodness-of-fit statistics revealed that the initial model built from the TAKS blueprint (Figure 15), is less than optimal for Group 1 (MLM$\chi^2_{[734]}$ = 1042.110; CFI=

0.845; RMSEA= 0.020). The advised cutoff value for the Comparative Fitness Index (CFI) to consider a model a good fit to the data is 0.95 (Byrne, 2012). In this case, the CFI of 0.845 indicates a poor fit of the model for Group 1. According to this initial result, the hypothesized model was modified using the Modification Index provided by MPlus, which identifies the misspecified parameters. Further specification of the modified model included four cross-loadings – items 6, 36, 39 and 40. Items 6 and 36 load on Factors 1 and 2, while items 39 and 40 load on Factors 1 and 3. Only items 6 and 36 were respecified, because the model did not converge when specifying the crossloadings for items 39 and 40. Item 1 was dropped as it was not statistically significant (Figure 16). Despite these modifications done to the model, the goodness-of-fit statistics revealed that although an improvement was achieved over the baseline model, the modified model for the 2009 Science TAKS is less than optimal for Group 1 ($MLM\chi^2_{[694]}$ = 947.669; CFI= 0.873; RMSEA= 0.018). Table 14 shows the standardized parameter estimates for the final TAKS model adjusted for Group 1. From Table 14, it can be noticed that items 6 and 36 are not a good fit to the model, as their parameter estimates are larger than 1, suggesting further revision of both items. Given these results, the models illustrated in Figures 15 and 16 are rejected.

High correlations between the four factors were found (with values between 0.814 and 1), suggesting that reducing the number of factors might yield to a more suitable model. Thus, further specification of this model is required, although for the purpose of this study, focusing on the equivalence of constructs across groups, it is sufficient to notice the lack of fit of the hypothesized TAKS baseline model.

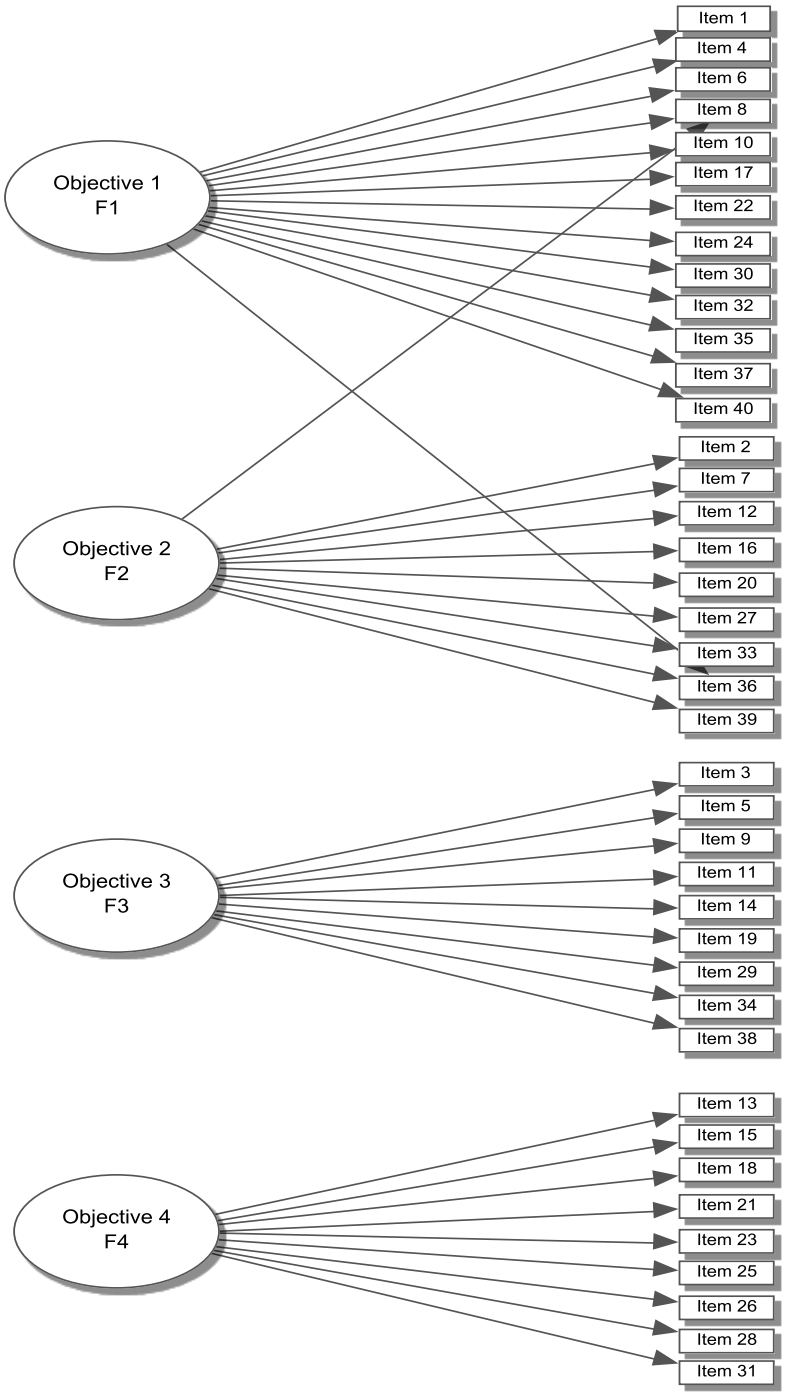


Figure 16: Final model of TAKS structure for Group 1

Factor	Item	Estimate	Standard Error	p
1	4	0.184	0.081	0.022
	6	1.827	0.577	0.002
	8	0.148	0.074	0.046
	10	0.543	0.074	0.000
	17	0.537	0.060	0.000
	22	0.365	0.054	0.000
	24	0.552	0.062	0.000
	30	0.455	0.069	0.000
	32	0.479	0.077	0.000
	35	0.410	0.073	0.000
	36	-0.549	0.196	0.005
	37	0.426	0.088	0.000
	40	0.639	0.050	0.000
2	2	0.383	0.081	0.000
	6	-1.502	0.588	0.011
	7	0.510	0.065	0.000
	12	0.266	0.094	0.005
	16	0.499	0.054	0.000
	20	0.647	0.064	0.000
	27	0.672	0.061	0.000
	33	0.566	0.070	0.000
	36	1.056	0.217	0.000
	39	0.559	0.093	0.000
3	3	0.270	0.069	0.000
	5	0.271	0.065	0.000
	9	0.410	0.063	0.000
	11	0.537	0.054	0.000
	14	0.552	0.061	0.000
	19	0.488	0.047	0.000
	29	0.472	0.066	0.000
	34	0.542	0.068	0.000
	38	0.145	0.078	0.002
4	13	0.452	0.057	0.000
	15	0.590	0.042	0.000
	18	0.469	0.045	0.000
	21	0.409	0.047	0.000
	23	0.523	0.047	0.000
	25	0.472	0.046	0.000
	26	0.504	0.048	0.000
	28	0.506	0.044	0.000
	31	0.437	0.060	0.000

Table 14: CFA Parameter estimates for the final TAKS model for Group 1

In addition, the Modification Index suggested that items 39 and 40 (see Figure 17) were misspecified. Both items load in two factors, but, when specifying the model for both cross loadings, the model did not converge. This, in addition to the low CFI, constitutes evidence of the fact that the items are not assessing the intended objectives for Group 1 students. Item 40 (Figure 17) intended to assess objective 1, more specifically that students "use critical thinking and scientific problem solving to make informed decisions" (TEA, 2009c). Pulling the actual item for a finer-grain analysis makes the misclassification more evident as the item does not require from students to make an informed decision based on critical thinking or problem solving.

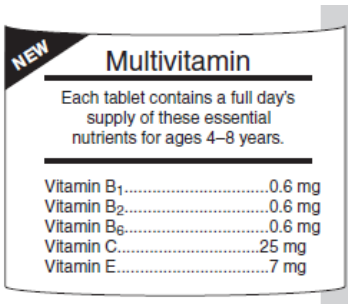
Item 39	Item 40
39 Crows are found throughout many parts of the world. They are black birds with excellent hearing. In the wild they live for six to seven years. Some crows in Japan open hard-shelled nuts by dropping them in front of moving cars. Which of these is a learned behavior? **A** Having excellent hearing **B** Living six to seven years **C** Dropping nuts in front of cars **D** Having black feathers	 **40** The picture shows the label for a new product. Based on the label, the reader can conclude that the tablets — **F** were made for children **G** have different flavors **H** are easy to chew **J** have a pleasant taste
Retrieved from http://www.tea.state.tx.us/student.assessment/taks/released-tests/archive/	

Figure 17: Items 39 and 40 taken from the released 2009 fifth grade Science TAKS

Baseline model for groups 2 (Hispanic, testlang English, non-LEP) and 4(Hispanic, testlang English, LEP)

The Goodness-of-fit statistics revealed that the initial model built from the TAKS blueprint (Figure 15) represents a good fit to the data for Group 2 ($MLM\chi^2_{[734]}$ = 819.356; CFI= 0.980; RMSEA= 0.011) and Group 4 ($MLM\chi^2_{[734]}$ = 805.124; CFI= 0.985; RMSEA= 0.010). Thus we fail to reject the model illustrated in Figure 15 for Groups 2 and 4. The model did not present cross-loadings. The MGCFA results indicate that the construct(s) assessed are similar for Groups 2 and 4, but are different than the one(s) assessed for Group 1. Tables 15 and 16 show the standardized parameter estimates for the hypothesized TAKS model adjusted for Groups 2 and 4 respectively. From both tables, it can be noticed that there are no misspecified items as it happened with items 6, 36, 39 and 40 for Group 1.

The results of the baseline models suggest important structural differences between the models that fit the data for each group, and constitute evidence of the difference between the construct (s) assessed for the three groups.

Factor	Item	Estimate	Standard Error	p
1	1	0.660	0.071	0.000
	4	0.573	0.068	0.000
	6	0.390	0.066	0.000
	8	0.477	0.065	0.000
	10	0.502	0.055	0.000
	17	0.378	0.051	0.000
	22	0.417	0.044	0.000
	24	0.624	0.041	0.000
	30	0.528	0.047	0.000
	32	0.654	0.050	0.000
	35	0.534	0.051	0.000
	37	0.610	0.058	0.000
	40	0.482	0.094	0.000
2	2	0.590	0.066	0.000
	7	0.448	0.050	0.000
	12	0.562	0.049	0.000
	16	0.404	0.045	0.000
	20	0.728	0.037	0.000
	27	0.489	0.054	0.000
	33	0.566	0.047	0.000
	36	0.518	0.060	0.000
	39	0.668	0.057	0.000
3	3	0.585	0.058	0.000
	5	0.500	0.064	0.000
	9	0.555	0.048	0.000
	11	0.440	0.045	0.000
	14	0.521	0.047	0.000
	19	0.514	0.037	0.000
	29	0.495	0.045	0.000
	34	0.411	0.057	0.000
	38	0.457	0.058	0.000
4	13	0.468	0.047	0.000
	15	0.593	0.038	0.000
	18	0.386	0.044	0.000
	21	0.354	0.045	0.000
	23	0.511	0.043	0.000
	25	0.392	0.044	0.000
	26	0.537	0.041	0.000
	28	0.507	0.042	0.000
	31	0.454	0.053	0.000

Table 15: CFA Parameter estimates for the hypothesized TAKS model for Group 2

Factor	Item	Estimate	Standard Error	p
	1	0.682	0.058	0.000
	4	0.628	0.045	0.000
	6	0.492	0.046	0.000
	8	0.428	0.058	0.000
	10	0.504	0.054	0.000
	17	0.349	0.049	0.000
1	22	0.331	0.044	0.000
	24	0.530	0.043	0.000
	30	0.524	0.043	0.000
	32	0.578	0.045	0.000
	35	0.511	0.047	0.000
	37	0.608	0.046	0.000
	40	0.650	0.067	0.000
	2	0.572	0.050	0.000
	7	0.477	0.051	0.000
	12	0.482	0.047	0.000
	16	0.457	0.042	0.000
2	20	0.650	0.039	0.000
	27	0.556	0.048	0.000
	33	0.543	0.042	0.000
	36	0.494	0.055	0.000
	39	0.660	0.045	0.000
	3	0.467	0.065	0.000
	5	0.510	0.057	0.000
	9	0.609	0.051	0.000
	11	0.422	0.050	0.000
3	14	0.511	0.045	0.000
	19	0.482	0.043	0.000
	29	0.451	0.046	0.000
	34	0.472	0.054	0.000
	38	0.543	0.069	0.000
	13	0.338	0.046	0.000
	15	0.634	0.036	0.000
	18	0.292	0.050	0.000
	21	0.307	0.046	0.000
4	23	0.436	0.046	0.000
	25	0.296	0.046	0.000
	26	0.531	0.041	0.000
	28	0.512	0.042	0.000
	31	0.473	0.048	0.000

Table 16: CFA Parameter estimates for the hypothesized TAKS model for Group 4

Based on the three analyses presented in this chapter, I can say that the functioning of the items is not stable across the groups of students identified for this study, and even across students according to their ability level. The results of the confirmatory factor analysis indicate that the constructs assessed by the English version of the 2009 Science TAKS also vary across ethnic and English proficiency groups. A finer analysis of the items and the way in which students respond is needed, to determine whether the constructs intended to be assessed match the actual knowledge elicited by the items.

Chapter V: Discussion and Conclusions

The conceptual model for this study, depicted in Figure 18, indicates the relevance of different sources of information that need to be considered to assess the validity of standardized tests for ELLs and the use of such evidence to interpret students' test scores. This is a cyclical and key process in understanding the validity of the differences in test scores between ELL and non-ELL populations.

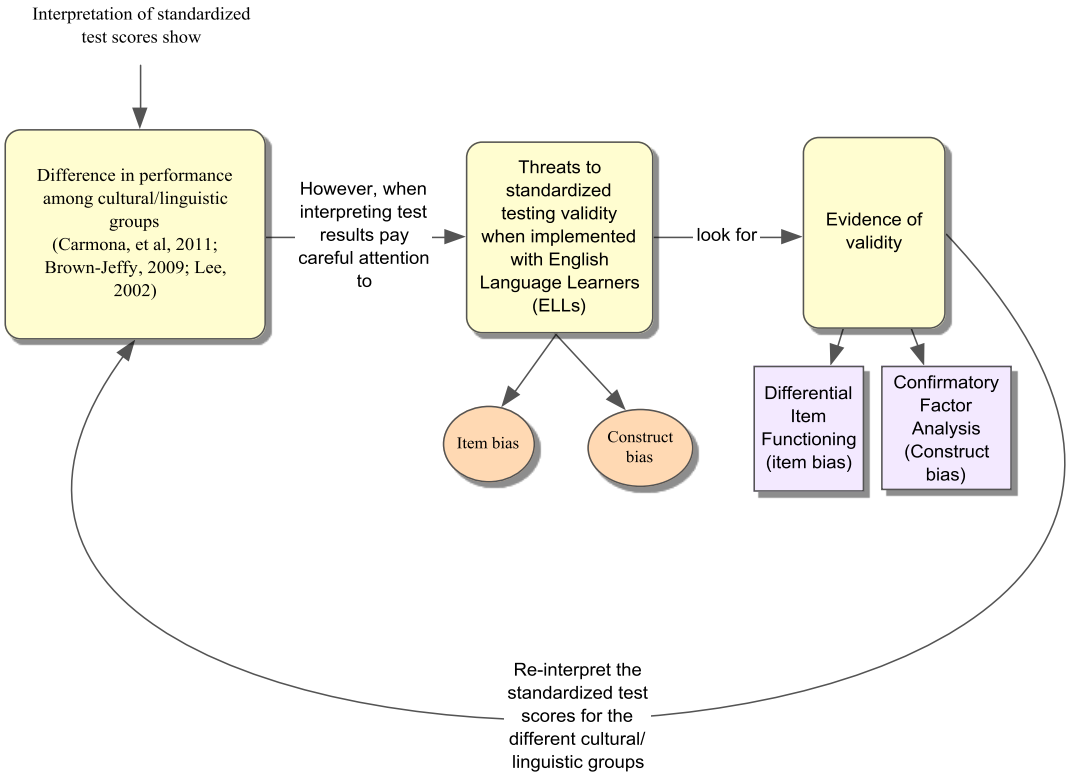


Figure 18: Conceptual framework

Even when most of the times such differences are considered as true differences in what students know, the results of this study presented in Chapter 4 show that these differences

are a function of student characteristics that interact differently with test items, creating inequitable assessments for different population groups. The high variability in functioning of the 2009 Science TAKS items, as well as the difference in constructs assessed across three of the four populations considered for the study became evident in this analysis. Such variability in the constructs measured is crucial in the interpretation of standardized test scores, if we consider that "standardized tests are administered, scored, and interpreted in a standard manner" (Reynolds, Livingston & Wilson, 2006, p.7).

The differences found in item functioning across groups matched on ability and constructs assessed by the 2009 fifth grade Science TAKS are evidence of the fact that the test is not assessing the same constructs for the four groups, and that the difficulty level of some items is not the same for all students. Thus assessing differences in students' science knowledge is difficult, as the sample of the knowledge that is being obtained from the four populations pertain to different constructs, at different levels of difficulty. According to van de Vijver and Poortinga (1997), the presence of nonuniform biased items –this is items that do not always advantage the same group, affect the construct equivalence of the test. Thus, the comparison of scores across cultures and languages are likely to yield invalid and inequitable results.

The purpose of the ANOVA analysis conducted for this study was to explore the differences in scores between the four groups considered for this study. The results of the Analysis of Variance reported in Chapter 4 are consistent with what has been reported in the literature regarding the achievement gaps existing between different ethnic, race and socioeconomic groups (Johnson, 2002, National Center for Education Statistics, 2011).

Pairwise comparisons showed that White, non-LEP students who answered the English version of the test score higher than Hispanic students, whether they are classified as LEP or non-LEP, and answered the English or Spanish version of the test. One of the issues of comparing achievement of linguistic and/or ethnic groups is controlling for variables that might contribute to such achievement differences, such as socioeconomic status. The grouping of students used for this study showed that ethnicity, test language, LEP classification and socioeconomic status are confounded variables.

Even when the ANOVA results reported in this study show statistically significant differences between the four student groups, such differences should not be considered as true differences in science knowledge assessed, until further revision of the items' psychometric properties, and studies of the differences/similarities in the content elicited by the items in students from the different groups, are conducted. Nevertheless, it is important to keep in mind that researchers have been advocating for the use of multiple measures to assess ELLs' content knowledge in order to provide more valid measures (Noble, et al., 2012; Valenzuela, 2002).

The ANOVA results presented in this dissertation also call to rethink the conceptualization of the achievement gap, depending on the evidence used to claim the existence of achievement differences among different student groups. If the term "achievement gap" is used to indicate achievement differences between ethnic/ linguistic/ cultural groups, then we need to ensure that the measures indeed refer to achievement and not to (for example) group characteristics like ethnicity, language, or socioeconomic status. At the classroom level, the fact that the differences in test scores between groups

does not reflect true differences in what students know or are able to do, call to re-consider the use of test scores to make instructional decisions based on differences in scores across student populations.

ASSESSING VALIDITY: ITEM AND CONSTRUCT BIAS

The concept of validity is comprised by different dimensions or sources of evidence that become especially important when interpreting test scores across diverse groups. When testing the knowledge of diverse groups simultaneously, validity goes in hand with fairness in what large-scale test scores should accomplish in providing comparable construct validity across groups (National Research Council [NRC], 2001). For this to happen, tests used to make inferences regarding different populations should not be biased towards a particular group or groups. In this study the presence of bias was analyzed mainly from the *psychometric* stance, and in less degree from a *linguistic* and *cultural* perspective.

From a *psychometric* stance, the test was analyzed at an item level to identify those items that might function differently across the four populations, and holistically to compare the constructs assessed across three of the four populations through a structural analysis.

At the *item level*, it can be said that some of the 2009 Science TAKS items function differently across the four groups considered for the analysis. This is, that a group of items is not psychometrically equivalent across the four groups in terms of their difficulty level, or the probability of endorsing the item according to students' ability level. The Mantel-Haenszel procedure revealed the presence of a large number of biased

items, and the ICCs depicted for some of the items are evidence of the existence of nonuniform bias. This is items do not always advantage the same group. The presence of nonuniform bias affects the equivalence of assessed constructs across groups. This creates validity issues, as not all students are being assessed in the same constructs, leading to a situation in which no comparisons can be conducted across different groups' knowledge.

The smallest number of items (6) with DIF was detected among Groups 1 and 2, which only differ in ethnicity. Thus, it can be said that the 2009 Science items function psychometrically similar across White and non-ELL Hispanic students. In contrast, a large number of DIF items were found between groups 1 and 4, and 2 and 4, where only group 4 is comprised of ELL students. This result supports previous research showing that students' native language influences importantly the way in which students' answer the items (Solano-Flores, Lara, Sexton, & Navarrete, 2001; Solano-Flores & Li, 2009; Solano-Flores & Trumbull, 2003). The influence of students' language in the item functioning indicates that students are not being fairly tested, and their English proficiency is interfering with students' ability to show what they know. Consequently, it can be said that English proficiency is a source of construct-irrelevant variance (Messick, 1995), as it introduces an aspect that is extraneous to the Science knowledge being assessed, making the items more difficult for some Hispanic LEP students.

LEP students are not the only population that is not being tested appropriately. The presence of items that do not differentiate between high and low performance students can yield to invalid comparison between students from the same Group. Popham (1999)

argues that one of the items' characteristics that allow making inferences about a students' status with respect to the mastery of certain content knowledge, is items' differentiation between low and high ability students. Thus, the test power to compare students relies on the discrimination index of the items. It is important to remember that construct-irrelevant easiness is also considered a source of invalidity (Messick, 1995).

At the *construct* level, evidence was found regarding the constructs being measured for each group, but also at how the items supported the hypothesized structure of the test. In relation to the number of constructs assessed for each group, the hypothesized model (Figure 15) for Group 1 was rejected, but not for Groups 2 and 3. This provides evidence that different constructs were assessed by the 2009 fifth grade Science TAKS for each group. Loevinger (1957) refers to structural validity as the "extent to which structural relations between test items parallel the structural relations of other manifestations of the trait being measured" (p. 661). Thus, as the internal structure of the test is different for each group, it can be concluded that the constructs they represent are different and in the case of Group 1, the structure of these set of items do not mirror the hypothesized content structure. The difference in constructs measured across groups, point to a problem with the internal consistency of the test. According to Loevinger (1957) the items selected to construct a test, should constitute a representative sample of the construct being assessed. In this case, the fact that some of the items intended to assess Objective 1 for Group 1 have small or negative loadings, and two of them load on other constructs, indicate that the content assessed by those items is not

clearly defined and is not aligned completely to the construct that is intended to be assessed.

The psychometric evidence collected through the DIF analyses and the Confirmatory Factor Analysis for Multiple Groups, point to sources of invalidity for the 2009 Science TAKS that should be considered for further construction of large-scale assessments, as comparing students from different cultural, ethnic and linguistic groups is mandated by the No Child Left Behind Act of 2001, and is being considered as a way to close the achievement gap between different student groups.

From a *cultural* and *linguistic* perspective, researchers have recognized the difficulties of creating equivalent assessments for students from different cultural and linguistic backgrounds (e.g. Lee & Buxton, 2010; Luykx et al., 2007; Martiniello, 2013; Moskovich, 2013; Solano-Flores & Nelson-Barber, 2001). In this case, it was noticed that the change of answer choices from the English to the Spanish version, changed the difficulty level of the item. Based on the results presented in Chapter IV I can say that the transadaptation process was not very successful, in terms of the possibility of generating items with contexts rooted in different cultures that are familiar to students, have the same psychometric properties and assess the same constructs across cultures. Research of other issues related to language and culture, were limited as the number of items common to the English and Spanish version was small.

This study was able to identify sources of invalidity of the 2009 Science TAKS, which have their origin, in the differential functioning of the items and structure of the tests. The analyses presented here provided evidence to show the difficulties of comparing

students from diverse cultural and linguistic backgrounds, indicating that the inferences made about students' knowledge using the 2009 Science TAKS are not be valid, and that even when sources of invalidity are attended to, at the cultural and linguistic level is hard to generate equivalent items and/or tests.

LIMITATIONS AND FUTURE RESEARCH DIRECTIONS

Two limitations are highlighted from this study. The first limitation is the characterization of the populations using characteristics such as socioeconomic status and Limited English Proficiency. These characteristics yield to populations that are highly heterogeneous, and more detailed research of the way in which these characteristics interact with students' responses to items is limited.

The second limitation is the small number of items that were common to the English and Spanish versions of the 2009 Science TAKS. The small number of common items limited the information available from the DIF and the Confirmatory Factor analyses. Even when most of the items used in the Spanish version were supposed to be a direct translation of the English items (TEA & Pearson, 2010), only 11 items were found to be a direct translation of the English version of the TAKS, and the other 29 items seem to have been originated in Spanish. TEA and Pearson need to provide more information about the Spanish version of the test in order to conduct further research, and continue to provide evidence to answer some of the key issues in ELL testing, including: (1) the language in which ELLs should be tested, and (2) how to make tests that allow making comparisons of students' knowledge between cultural and linguistically diverse populations.

References

Abedi, J. (2002). Standardized achievement tests and English Language Learners: Psychometric issues. *Educational Assessment, 8*(3), 231-257.

Abedi, J. (2011). Assessing English Language Learners: Critical issues. In M. d. R. Basterra, E. Trumbull & G. Solano-Flores (Eds.), *Cultural Validity in Assessment*. New York: Routledge.

Abedi, J., Courtney, M., Mirocha, J., Leon, S., & Goldberg, J. (2005). *Language accommodations for English language learners in large-scale assessments: Bilingual dictionaries and linguistic modification.* Los Angeles, CA: Center for the Study of Evaluation. National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Hofstetter, C., & Lord, C. (2004). Assessment accommodations for English Language Learners: Implications for policy based empirical research. *Review of Educational Research, 74*(1), 1-28.

Abedi, J., Leon, S., & Mirocha, J. (2003). *Impact of student language background on content-based performance: Analyses of extant data*. Los Angeles, CA: Center for the Study of Evaluation. National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Lord, C., Kim, C., & Miyoshi, J. (2001). *The effects of accommodations on the assessment of Limited English Proficiency (LEP) students in the National Assessment of Educational Progress (NAEP)*. Los Angeles, CA: Center for the

Study of Evaluation. National Center for Research on Evaluation, Standards and Student Testing.

Brislin, R. W., & Freimanis, C. (2001). Back-translation: A tool for cross-cultural research. In C. Sin-Wai & D. E. Pollard (Eds.), *An Encyclopedia of Translation*. Hong Kong: The Chinese University Press.

Brown, T. A. (2003). Confirmatory factor analysis of the Penn State Worry questionnaire: Multiple factors or method effects. *Behavior Research and Therapy, 41,* 1411-1426.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: The Guildford Press.

Butler, F. A., & Stevens, R. (2001). Standardized assessment of the content knowledge of English language learners K-12: current trends and old dilemmas. *Language Testing, 18*(4), 409-427.

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema, 20*(4), 872-882.

Byrne, B. (2012). *Structural Equation Modeling with Mplus: Basic concepts, applications, and programming*. New York, NY: Routledge.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Carmona, G., Krause, G., Monroy, M., Lima, C., Avila, M. A., & Ekmecki, A. (2011). *A Longitudinal Study to Investigate Changes in Students' Mathematics Scores in Texas*. Paper presented at the AERA 2011 Annual Meeting.

Darling-Hammond, L., Barron, B., Pearson, P. D., Schoenfeld, A. H., Stage, E. K., Zimmerman, T. D., et al. (2008). *Powerful Learning: What we know about teaching for understanding*. San Francisco, CA: Jossey-Bass.

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York, NY: The Guildford Press.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, New Jersey: Lawrence Erlbaum.

Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French Versions of TIMSS. *International Journal of Testing, 5*(1), 23-35.

Gamst, G., Meyers, L. S., & Guarino, A. J. (2008). *Analysis of variance designs: A conceptual and computational approach with SPSS and SAS*. New York: Cambridge University Press.

Garcia, G. E., & Pearson, P. D. (1994). Assessment and Diversity. *Review of Research in Education, 20*, 337-391.

Geisinger, K. F. (1994). Cross-Cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6*(4), 304-312.

Grisay, A. (2003). Translation procedures in OECD-PISA international assessment. *Language Testing, 20*(2), 225-240.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education, 15*(3), 309-333.

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3-38). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hambleton, R., & Rodgers, J. H. (1995). Item bias review [Electronic Version]. *Practical Assessment, Research & Evaluation*, *4*. Retrieved October 24, 2012, from http://pareonline.net/getvn.asp?v=4&n=6

Johnson, R. S. (2002). *Using data to close the achievement gap: How to measure equity in our schools* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Kazemi, E. (2002). Exploring test performance in mathematics: the questions children's answers raise. *Journal of Mathematical Behavior, 21*(2), 203-224.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). What do test scores in Texas tell us? [Electronic Version]. *Education Policy Analysis Archives*, *8*, 1-22, from http://epaa.asu.edu/ojs/article/view/440/563

Kopriva, R. (2008). *Improving Testing For English Language Learners: A Comprehensive Approach to Designing, Building, Implementing & Interpreting Better Academic Assessments*. from http://UTXA.eblib.com/patron/FullRecord.aspx?p=330980

Lacelle-Peterson, M. W., & Rivera, C. (1994). Is it real for all kids? A framework for equitable assessment policies for English Language Learners. *Harvard Educational Review, 64*(1), 55-75.

Ladson-Billings, G. (2006). From the Achievement Gap to the Education Debt: Understanding achievement in U.S. Schools *Educational Researcher, 35*(3), 3-12.

Laosa, L. M. (1977). Nonbiased assessment of children's abilities: Historical antecedents and current issues. In T. Oakland (Ed.), *Psychological and Educational Assessment of Minority Children*. New York: Brunner-Routledge.

Lee, J. (2002). Racial and Ethnic Achievement Gap trends: Reversing the progress toward equity? *Educational Researcher, 31*(3), 3-12.

Lee, O., & Buxton, C. A. (2010). *Diversity and equity in Science Education: Research, Policy and Practice*. New York: Teachers College Press.

Lee, O., & Luykx, A. (2006). *Science education and student diversity: Synthesis and research agenda*. New York, NY: Cambridge University Press.

Linton, T. H., & Kester, D. (2003). Exploring the achievement gap between white and minority students in Texas: A comparison of the 1996 and 2000 NAEP and TAAS eight grade Mathematics test results [Electronic Version]. *Education Policy Analysis Archives, 11,* from http://epaa.asu.edu/ojs/index.php/epaa/article/view/238?lang=pt

Loevinger, J. (1957). OBJECTIVE TESTS AS INSTRUMENTS OF PSYCHOLOGICAL THEORY: Monograph Supplement 9. *Psychological reports, 3*(3), 635-694.

Luykx, A., Lee, O., Mahotiere, M., Lester, B., Hart, J., & Deaktor, R. (2007). Cultural and Home Language Influence on Children's Responses to Science Assessment. *Teachers College Record, 109*(4), 897-926.

Magis, M., Beland, S., & Raiche, G. (2013). difR [Computer software]. Retrieved from http://cran.r-project.org/web/packages/difR/index.html

Martiniello, M. (2013). *Diversity and Equity: Assessment challenges and examples for English Learners*. Paper presented at the CIME MSRI.

Martiniello, M., & Wolf, M. K. (2012). Exploring ELLs' understanding of word problems in Mathematics assessments - The role of text complexity and student background knowledge. In S. Celedón-Pattichis & N. G. Ramirez (Eds.), *Beyond good teaching: Advancing Mathematics education for ELLs*. Reston, VA: National Council of Teachers of Mathematics.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1995). Validity of Psychological Assessment. *American Psychologist, 50*, 741-749.

Moskovich, J. (2013). *Recommendations for formative Mathematics assessment for English learners*. Paper presented at the CIME MSRI.

National Center for Education Statistics. (2011). *Achievement Gaps: How Hispanic and White students in public schools perform in Mathematics and Reading on the National Assessment of Educational Progress* (NCES 2011-459). Retrieved from http://nces.ed.gov/nationsreportcard/pdf/studies/2011459.pdf

National Research Council. (2001). *Knowing what students know: The science and Design of Educational Assessment*. Washington, DC: National Academies Press.

National Research Council. (2007). *Taking Science to School: Learning and Teaching Science in Grades K-8*. Washington, DC: The National Academies Press.

No Child Left Behind (NCLB) Act of 2001, 20 U.S.C.A. § 6301 *et seq.* (West 2003)

Noble, T., Suarez, C., Rosebery, A., O'Connor, M. C., Warren, B., & Hudicourt-Barnes, J. (2012). "I never thought of it as freezing": How students answer questions on large-scale science tests and what they know about science. *Journal of Research in Science Teaching, 49*(6), 778-803.

Popham, W. J. (1999). Why standardized tests don't measure educational quality. *Educational Leadership, 56*(6), 8-15.

Reynolds, C. R., Livingston, R. B., & Wilson, V. (2006). *Measurement and Assessment in Education*: Pearson Education, Inc. .

Santel-Parke, C., & Cai, J. (1997). Does the task truly measure what was intended? *Mathematics Teaching in the Middle School, 3*, 74-82.

Singham, M. (2003). The achievement gap: Myths and reality. *Phi Delta Kappan, 84*(8), 586-591.

Sireci, S. G., Patsula, L., & Hambleton, R. K. (2005). Statistical Methods for Identifying Flaws in the Test Adaptation Process. In R. K. Hambleton, P. F. Merenda & C. D. Spielberger (Eds.), *Adapting Educational and Psychological Tests for Cross-Cultural Assessment* (pp. 93-116). Hillsdale, NJ: Lawrence Erlbaum Associates.

Solano-Flores, G. (2011). Assessing the Cultural Validity of Assessment Practices. In M. d. R. Basterra, E. Trumbull & G. Solano-Flores (Eds.), *Cultural Validity in Assessment*. New York: Routledge.

Solano-Flores, G., Backhoff, E., & Contreras-Niño, L. A. (2009). Theory of Test Translation Error. *International Journal of Testing, 9*(2), 78-91.

Solano-Flores, G., & Gustafson, M. (2013). Academic assessment of English Language Learners: A critical, probabilistic, systemic view. In M. Simon, K. Ercikan & M. Rousseau (Eds.), *Improving large-scale assessment in education: Theory, issues, and practice* (pp. 87-109). New York: Routledge.

Solano-Flores, G., Lara, J., Sexton, U., & Navarrete, C. (2001). *Testing English Language Learners: A Sampler of Student Responses to Science and Mathematics Test Items*. Washington, DC: Council of Chief State School Officers.

Solano-Flores, G., & Li, M. (2009). Language Variation and Score Variation in the Testing of English Language Learners, Native Spanish Speakers. *Educational Assessment, 14*(3-4), 180-194.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the Cultural Validity of Science Assessments. *Journal of Research in Science Teaching, 38*(5), 553-573.

Solano-Flores, G., & Trumbull, E. (2003). Examining Language in Context: The Need for New Research and Practice Paradigms in the Testing of English-Language Learners. *Educational Researcher, 32*(2), 3-13.

Solano-Flores, G., & Trumbull, E. (2008). In what language should English language learners be tested? In R. J. Kopriva (Ed.), *Improving testing for English language learners*. New York, NY: Routledge.

Texas Education Agency. (2004). TAKS Information Booklet: Elementary Science Grade 5.

Texas Education Agency. (2009a). *TAKS Data file format with student item analysis*. Retrieved from http://www.tea.state.tx.us/student.assessment/datafileformats/

Texas Education Agency. (2009b). TAKS Statewide Summary Reports 2008-2009. Retrieved from http://www.tea.state.tx.us/student.assessment/taks/rpt/sum/yr09/

Texas Education Agency. (2009c). Texas Assessment of Knowledge and Skills –Answer Key. Retrieved from http://www.tea.state.tx.us/student.assessment/released-tests/

Texas Education Agency. (2010). *Enrollment in Texas Public Schools 2009-10*. (Document No.GE11 601 01)Austin, TX: Author.

Texas Education Agency. (2012). *Enrollment in Texas Public Schools 2011-12*. (Document No. GE13 601 02) Austin, TX: Author.

Texas Education Agency, & Pearson. (2010). Technical Digest for the Academic Year 2009-2010. Retrieved from http://www.tea.state.tx.us/student.assessment /techdigest/yr0910/

Texas Education Code, §29.052 (West, 1991).

Trumbull, E., & Solano-Flores, G. (2011). The Role of Language in Assessment. In M. d. R. Basterra, E. Trumbull & G. Solano-Flores (Eds.), *Cultural Validity in Assessment*. New York: Routledge.

Turkan, S., & Liu, O. L. (2012). Differential performance by English Language Learners on an inquiry-based Science assessment. *International Journal of Science Education, 34*(15), 2343-2369.

U.S. Department of Education. (2006*).* Building partnerships to help English Language Learners Fact Sheet. Retrieved from http://www2.ed.gov/nclb/methods/english/lepfactsheet.pdf

U.S. Department of Education, Institute of Education Sciences, & National Center for Education Evaluation and Regional Assistance. (2009) *What Works Clearinghouse: Procedures and Standards Handbook* (Version 2.1). Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v2_1_standards_handbook.pdf

Valenzuela, A. (2002). High-Stakes testing and U.S. - Mexican youth in Texas: The case for multiple compensatory criteria in assessment. *Harvard Journal of Hispanic Policy*, *14*, 97-116.

van de Vijver, F., & Leung, K. (1997). *Methods and Data Analysis for cross-cultural research*. Thousand Oaks, CA: Sage.

van de Vijver, F., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Verschaffel, L., De Corte, E., & Lasure, S. (1994). Realistic considerations in mathematical modeling of school arithmetic word problems. *Learning and Instruction, 4*, 273-294.

Wiliam, D. (2008). International comparisons and sensitivity to instruction. *Assessment in Education, 15*(3), 253-257.

Wolf, M. K., & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. *Educational Assessment, 14*, 139-159.

Zucker, S., Miska, M., Alaniz, L. G., & Guzmán, L. (2005). *Transadaptation: Publishing assessments in world languages*. San Antonio, TX: Pearson Education.

Vita

Cynthia Esperanza Lima Gonzalez was born in Mexico City. After completing her work at Colegio Martinak, in 1996, she entered Universidad Nacional Autónoma de México, México City. She received the degree of Bachelor of Physics from Universidad Nacional Autónoma de México in September, 2004. In September 2007, she entered the Graduate School at the University of Texas at Austin.

Permanent Address: 263 Spring Drive
Kyle, Texas 78640

This thesis was typed by Cynthia Esperanza Lima Gonzalez.